UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Identificación de historia de las noticias en medios de comunicación online

Andrés Calvente Rodríguez
Tutor: Francisco Jurado Monroy

Julio 2020

Identificación de historia de las noticias en medios de comunicación online

AUTOR: Andrés Calvente Rodríguez
TUTOR: Francisco Jurado Monroy

Dpto. Ingeniería informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio de 2020

Resumen

En la actualidad se está viviendo un auge en la era de la información y de las nuevas tecnologías. La gente tiene muchas alternativas para conocer los sucesos que actualmente se están produciendo y alguna información puede ser falsa o no estar totalmente contrastada. Por ello, este **Trabajo de Fin de Grado propone crear una aplicación que, en base a una noticia seleccionada de un periódico, extraiga un historial cronológico de noticias para saber qué sucesos se han producido a lo largo del tiempo.** Esto permitiría a las personas conocer mejor el desarrollo de un evento y poder enfrentarse mejor a un posible debate o a una nueva noticia emitida por un canal no convencional.

Para abordar este objetivo se han tenido que hacer dos desarrollos, un módulo de extracción de noticias y otro módulo que construirá el historial de estas.

El primer módulo es esencial ya que el proyecto parte sin datos sobre las noticias que se han publicado durante los últimos años y son necesarias para poder compararlas entre sí y construir un historial. Para conseguir dicha información se han seleccionado **5 diarios de ámbito nacional**.

Una vez se tenga un conjunto de datos con información sobre las noticias, el segundo módulo calculará la similitud entre ellas haciendo uso de los atributos que las componen, siendo algunos de ellos el **titular**, las **palabras clave** o el **cuerpo**. En primer lugar, los textos serán tratados para obtener representaciones que permitan compararlos con la noticia seleccionada, obteniendo aquellos que traten sobre la misma temática. Finalmente, se calculará el orden cronológico de las noticias resultantes, de modo que el usuario pueda documentarse sobre el suceso seleccionado.

Palabras clave

Búsqueda y recuperación de información, Crawling Web, Scrapping Web, Procesamiento del Lenguaje Natural, Algoritmos de Similitud, Periodismo Digital.

Abstract

Nowadays, there is a boom in the information age and new technologies. People have many alternatives to know the events that are currently taking place and some information may be false or not fully contrasted. Therefore, this **Bachelor Thesis proposes to create an application that, based on a selected news item from a newspaper, extracts a chronological history of news to find out what events have occurred over time**. This will allow people to better understand the development of an event and to better face a possible debate or new news broadcast by an unconventional channel.

To achieve this objective, two developments have been made, a news extraction module and another module that will build the history of these news.

The first module is essential since the project starts without data of the news that have been published in recent years and is necessary to be able to compare each other and build the history. To obtain this information, **5 Spanish newspapers have been selected**.

Once the information has been retrieved, the second module will calculate the similarity between them using the attributes that make up a news item, some of them being the **headline**, the **keywords**, or the **body**. Firstly, the texts will be treated to obtain representations that allow them to be compared with the selected news, obtaining those that deal with the same subject. Finally, the chronological order of the resulting news will be calculated, so that the user can document the selected event.

Keywords

Information Search and Retrieval, Web Crawling, Web Scrapping, Natural Language Processing, Similarity Algorithms, Digital Journalism.

Agradecimientos

En primer lugar, agradéceselo a mis padres. Ellos han sido los que han vivido conmigo todos los momentos, tanto buenos como malos dentro y fuera del ámbito estudiantil. Ellos han sabido escucharme y aconsejarme, animándome a sobreponerme a los problemas y haciéndome ver la recompensa que hay detrás de ellos. Gracias por todo el esfuerzo que habéis hecho para que yo haya cursado este grado, que es lo que me ha formado como persona y me ha dado un objetivo en la vida. También quiero agradecer al resto de mi familia, tíos, primos y abuelos, por apoyarme y estar pendientes de cómo iba evolucionando el proyecto. Ellos me han inculcado los valores que ahora mismo me hacen ser la persona que soy y estoy feliz de serlo.

Agradecer especialmente a mi tutor, Francisco Jurado, por permitirme realizar este proyecto con él. Gracias por guiarme y estar dispuesto en todo momento a resolver todas las dudas que han ido surgiendo a lo largo del trabajo. Me alegro de que me hayas ayudado a desarrollar este proyecto ya que me ha permitido emplear muchos de los conocimientos que he aprendido en estos cinco años de carrera.

Agradecer a todos los profesores de la universidad que me han dado clase, ya que sin ellos no habría tenido los conocimientos suficientes para haber desarrollado este TFG y convertirme en el buen ingeniero que espero ser. También agradecer a todos mis profesores de primaria y secundaria por enseñarme a que aprender es una tarea que nunca acaba.

Por último y no menos importante, gracias a mis amigos. Vosotros sois los que me habéis hecho vivir las mejores experiencias, los mejores viajes, las mejores fiestas, las mejores risas. Muchas gracias a todos los que estuvisteis, estáis y estaréis compartiendo vuestra vida con la mía. No puedo estar más orgullo de lo que me habéis aportado como persona. Gracias a mi familia de vecinos por darme las mejores escapadas: Paula D., Paula M., Sara, Víctor. Gracias a toda la gente que he conocido en la Universidad, la nueva familia que he hecho, por haberme dado los mejores años de mi vida: Tomás, Carlos, Sergio, Nathan, Expo, Laura, Jesús, María, Juan, Blanca, Alberto, Miguel, Alfonso, Álvaro, Guille, Jaime, Raúl, Inés, Edu, Mario. Gracias a todos mis amigos de Alcobendas, que no importa cuánto tiempo pase, siempre nos sentiremos como el primer día: Paula, Carlos, María, Fran, Pablo, Maya, Raúl, Eric, Nacho, Alex, Sergio, Alberto, Daniel. Gracias a los camareros de la facultad, qué entre café y café, siempre tenían tiempo de escucharte y darte ánimos para seguir adelante: Diego, Lorenzo. A toda la gente que he conocido en mis trabajos, gracias por enseñarme el ámbito laboral y por haberme dado la oportunidad de aprender con vosotros: David, Vicente, Mariano, Gerardo, Nacho, David, Alejandro, Mauri, Pablo, Fers, Marta, Ros, Alejandra, Satur, Álvaro. Y gracias a todos los que no he nombrado, pero se han cruzado en mi vida y me han apoyado a seguir adelante con mis decisiones. Muchas gracias a todos por formar parte de ella.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

En la actualidad se está viviendo un auge en la era de la información y de las nuevas tecnologías. La gente dispone de muchas fuentes de información y un rápido acceso a todas ellas. Cualquiera que tenga acceso a una conexión a internet y un teléfono móvil puede resolver cualquier tipo de duda que se le genere, ¿qué tiempo hace hoy?, ¿cuál es la última noticia sobre el gobierno de nuestro país?, ¿cómo está la bolsa?, ¿qué conflictos mundiales se están produciendo ahora y por qué surgieron? Todas estas preguntas pueden ser resueltas en cuestión de minutos.

Aunque este es un gran avance para toda la humanidad, con tanta información al alcance de la mano, siempre está la posibilidad de que alguna información sea errónea o provenga de fuentes poco fiables. Por ello, la gente está expuesta a noticias falsas, que en la actualidad son conocidas con el término ***fakenews***[1], bulos que fluyen por las aplicaciones de mensajería y otras redes sociales. Por tanto, es difícil que solo con nuestro conocimiento podamos saber si una noticia es falsa o cierta.

Con este trabajo se pretende desarrollar un software con el cual se pueda conocer el historial de sucesos que se han ido publicando sobre una noticia en algunos medios de comunicación, más concretamente en la prensa. Se sabe que en cualquier redacción de un periódico, ya sea nacional o internacional de renombre, tienen **Documentalistas**[2], los cuales se dedican a atribuir una serie de etiquetas a una noticia para definirla. En base a estas etiquetas, los usuarios buscan las noticias con esa temática, pero en este proyecto se pretende estudiar otros modos de búsqueda utilizando la propia información de las noticias para generar resultados mejores al actual.

Si se lograse este objetivo en poco tiempo, muchas noticias falsas serían rápidamente detectadas por su incoherencia con el resto del historial de la noticia a la que hace referencia, pudiendo así identificar los bulos y evitar su rápida propagación.

1.2 Objetivos

Con la motivación anterior, los objetivos que se han definido para este proyecto son:

- **Desarrollar un programa que pueda crear el historial de una noticia**. Objetivo principal del proyecto con el cual se podrá obtener una cadena de noticias anteriores a una noticia seleccionada por el usuario, siendo todas estas relevantes para conocer los inicios de un suceso y cómo se han ido desarrollando los acontecimientos según avanzan los días.
- **Calcular la similitud entre noticias**. Para poder crear un historial de una noticia se tendrá que poder conocer de alguna manera si una noticia se parece a otra, de este modo solo formarán parte del historial las noticias relacionadas. Esto se realizará utilizando ciertas funciones matemáticas que son capaces de determinar la similitud entre textos.
- **Crear un módulo de tratamiento de textos**. En los textos que se publican, ya sean artículos científicos, documentaciones técnicas o, lo que atañe a este proyecto, noticias publicadas en un periódico, se utiliza un lenguaje rico en vocabulario y gramática, más incluso en una lengua como la nuestra. Por ello, es necesario desarrollar una funcionalidad que transforme los textos de las noticias en modelos o representaciones más sencillas, para poder llevar a cabo el cálculo de la similitud entre textos.

- **Crear nuestra propia colección de noticias**. Para poder crear un historial se necesita tener un cúmulo de noticias que abarque desde la fecha en el pasado que se precise, hasta la fecha de publicación de la noticia con la cual se quiera trabajar en el presente. Como este trabajo se centra en el histórico de noticias publicadas en periódicos y hay mucha variedad de editoriales y mucha diferencia entre las noticias que se publican en periódicos nacionales e internacionales, se ha decidido trabajar con cinco periódicos nacionales, siendo estos: **El País**[3], **El Mundo**[4], **20 Minutos**[5], **El Confidencial**[6] y **El Marca**[7].
- **Crear un crawler**[8] **que sea capaz de navegar por las diferentes webs de los periódicos**. Si se quiere crear un conjunto de datos de noticias de forma automática, necesitamos por ende una forma de poder recorrer todas las webs de una forma eficiente, localizando así todas las noticias para su futuro análisis.
- **Crear un scrapper**[9] **que sea capaz de extraer la información importante de una noticia**. Al igual que hemos planteado el crawler, también es necesario una herramienta que automatice el proceso de extracción de información de una noticia. Si bien podemos ver que los componentes de una noticia pueden ser muchos, nosotros hemos propuesto una serie de atributos con los que vemos suficiente la definición de una noticia. Estos serían: **Título**, **keywords**, **resumen de la noticia**, **autor/es**, **localización/es**, **fecha de publicación**, **pie de foto**, **firma del pie de la foto**, **cuerpo de la noticia** y **tags**. La diferencia que hay entre **keywords** y **tags** en una noticia es que la primera hace referencia a las palabras más utilizadas en un texto y la segunda son las etiquetas que la editorial del periódico ha decidido poner a la noticia para encasillarla en un grupo de temas.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1. Introducción:** Relata un análisis de la situación actual de las tecnologías de información, como los usuarios hacen uso de ellas y como les afectan en sus vidas. También se citan y describen los objetivos que planteamos resolver con el desarrollo de este proyecto.
- **Capítulo 2. Estado del arte:** Define conceptos de forma teórica que luego se implementarán en la parte práctica. Además, hará referencia a otros proyectos relacionados al que se ha desarrollado.
- **Capítulo 3. Análisis y Diseño:** En el apartado de análisis se definen los requisitos funcionales y no funcionales del proyecto. En el apartado de diseño se explica cómo se ha estructurado el proyecto para alcanzar los objetivos.
- **Capítulo 4. Desarrollo:** Se citan algunas decisiones importantes que se han tomado, las cuales repercutirán a la creación de las funcionalidades del proyecto. Además, se hace una explicación extensa sobre las funcionalidades creadas y como se ejecutan para obtener los resultados que necesitamos.
- **Capítulo 5. Pruebas y resultados:** Expone y discute las pruebas que se han hecho y los resultados obtenidos en ellas para evidenciar que el trabajo realizado obtiene los resultados esperados.
- **Capítulo 6. Conclusiones y trabajo futuro:** Hace un breve resumen de las ideas finales que aporta el proyecto y las tareas que pueden realizarse a posteriori para dotar de mejoras al trabajo realizado.

2 Estado del arte

En este capítulo se abordarán los conceptos que han sido necesarios estudiar de forma teórica para su posterior implementación. En particular se tratará el proceso de extracción de información de la web mediante técnicas de scrapping, las aproximaciones que existen de cara a realizar el proceso de recorrido de la web mediante crawling, las medidas de similitud entre textos, y algunos de los modelos de representación del mismo para poder computar dichas medidas.
Asimismo, se hará referencia otros proyectos relacionados al que se ha desarrollado a lo largo del presente TFG.

2.1 Scrapping

2.1.1 Definición

El Scrapping o Web Scrapping es la tarea por la cual se analiza una página Web para obtener cierta información de ella. El Scrapping permite la recolección de información del texto semi-estructurado en la Web, siendo muy útil para las grandes empresas, ofreciendo varios beneficios[10] como por ejemplo:

- Una rápida recolección de información sobre las opiniones de los usuarios en cuanto a un producto o sobre las decisiones que se están tomando.
- Permite analizar otras empresas o marcas, viendo cómo se pueden optimizar los precios de los productos para aumentar la competencia.
- Permite anticiparse a los productos o tendencias que se van a poner de moda, lo que se conoce como *cool hunting* o caza de tendencias. Con este tipo de información, una empresa podría llevarse un gran beneficio si la moda que se ha descubierto tiene un gran impacto y perdura en el tiempo.

Para entender mejor el concepto pongamos un breve ejemplo. Imagine que se tiene interés por extraer la información sobre los productos de un supermercado. Con obtener el nombre del producto y su precio es suficiente. La página de este supermercado lista todos sus productos en su Web, poniendo el nombre de cada producto con su precio, además de otra información como la marca del producto o la puntuación que los usuarios han dado al producto. Se podría ir a la página Web de este supermercado e ir copiando cada producto junto a su precio en una base de datos. Esto sería muy laborioso, imagine que más tarde es interesante agregar a cada producto la marca del fabricante o la puntuación de los usuarios, se debería de recorrer de nuevo el listado de productos, algo muy ineficiente. Por ello, proponiendo la solución de implementar un Scrapper, se podría guardar la Web del listado de productos, que es una página en HTML, y mediante reglas, adquirir los datos que interesen de cada producto, automatizando el trabajo y solucionando el problema.

2.1.2 Extracción de información de una Web

Sabiendo que una página Web está compuesta por código HTML, ahora la pregunta que surge es cómo se puede extraer automáticamente información de estos códigos. Pues bien, a continuación se exponen dos formas de localización de elementos en una página HTML que usan en la actualidad:

- **CSS Selectors**[11]

Esta herramienta es la utilizada para crear las fichas de estilos de las páginas Web, lo que se conoce comúnmente como ficheros CSS de estilos. En este lenguaje, se asigna a un tag de HTML unas características de estilo con el fin de proporcionar al elemento la apariencia que

se desee. Estas características pueden ser color del texto, altura, anchura, posición dentro de la Web, etc. La forma en la que se escriben los selectores de CSS es la siguiente:

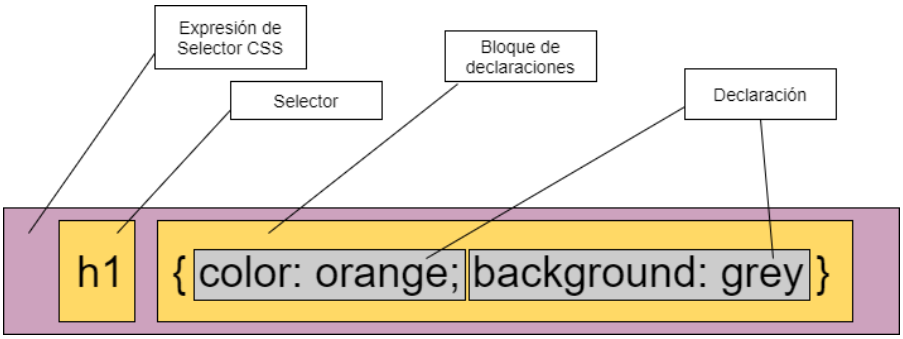

Figura 2-1: Ejemplo de selector CSS

Volviendo al tema del Scrapping, se pueden utilizar estos selectores para identificar elementos dentro del **DOM** de la página, siguiendo unas expresiones establecidas para encontrar los datos que se deseen. Esto es una ventaja, debido a que si el programador que está diseñando el scrapper está familiarizado con escribir hojas de estilo en CSS, no le resultaría muy complicado realizar las distintas expresiones para recolectar los datos.
Las desventajas que tienen los selectores de CSS es que, en páginas con mucho contenido HTML, las expresiones que se deben utilizar para llegar a los datos requeridos serán muy complejas. Esto es debido a que se debe de hacer uso de los hermanos/hijos de los diferentes tags para llegar a la información, haciendo bastante difícil la explicación de estas.

- **XPath**[12]

Nombre que viene de la forma en la que se crean las expresiones, como las **rutas (path)** de una URI, y al tipo de documentos al que van dirigidos, **XML**, es una herramienta pensada para la localización de elementos dentro de este tipo de documentos. Las expresiones **XPath** también se pueden utilizar en las Webs HTML.
Aunque comparte también la dificultad con los selectores CSS en cuanto a creación de expresiones, estas son algo más sencillas debido a la manera en la que se escriben, que como hemos dicho anteriormente, es como definir una ruta desde un elemento principal hasta el elemento que contiene el dato que deseamos. Un ejemplo sería el siguiente:

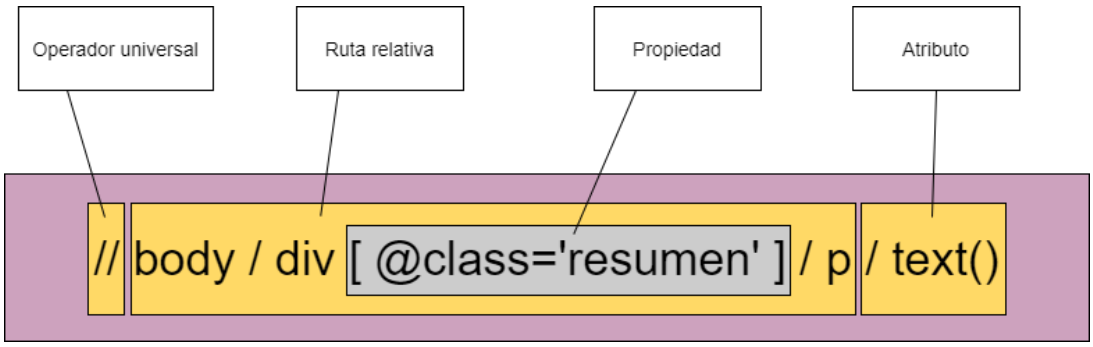

Figura 2-2: Ejemplo de expresión XPath

Una ventaja de esta herramienta respecto a los selectores CSS, es que estando en un nodo o elemento dentro del DOM, podemos volver al tag padre sí lo deseamos. Esta ventaja no cambia mucho la forma de creación de las expresiones, debido a que es más correcto poner la ruta directa a la información partiendo de un elemento sin tener que volver hacia atrás, pero aporta un punto de versatilidad a la herramienta que puede ser útil en algunos casos.

2.1.3 Tipos de implementación y herramientas

Aunque hay muchas formas de implementar este tipo de programas y herramientas que facilitan dicha implementación, se citarán las tres alternativas más usadas actualmente:

- **Scrapy**[13]

Framework open source en Python que, como se verá, se utilizará en este proyecto. Gracias al empleo de patrones de diseño software orientados a la generación especializada de este tipo de herramientas, Scrapy ofrece muchas ventajas, la primera de ellas es que la curva de aprendizaje de la herramienta es rápida, incluso teniendo tanto potencial. A parte, Scrapy ofrece también la posibilidad de hacer *crawling*, concepto que será definido más adelante, pero en avance, es la manera de navegar automáticamente por las páginas de una Web.

- **Beautiful Soup**[14]

Librería Python que extrae información de páginas HTML y XML. Beautiful Soup es una excelente herramienta para examinar el DOM de una página HTML y extraer datos de ellas. Su uso es muy intuitivo debido al nombre de sus funciones, además de que el objeto principal de la librería tiene como atributo toda la información sobre cada tag examinado en el fichero Web.

- **Selenium**[15]

Herramienta cuyo punto fuerte es proporcionar un emulador de un navegador con el cual se puede simular la interacción de un usuario normal con la Web. Esto puede ser interesante en algunos escenarios donde para acceder a la información requerida, primero se ha de pasar por un *Login* o una zona donde sea necesaria la interacción del usuario.

2.2 Crawling

2.2.1 Definición

Crawling se define como la acción de navegar entre las distintas páginas de la Web o de un dominio determinado, partiendo desde una página denominada **semilla** y con el fin de recopilar todas las páginas que componen ese dominio, pudiendo extraer la información que contienen. El Crawling es muy utilizado a la hora de identificar dónde se encuentra información, elemento fundamental para crear el índice de un motor de búsqueda. El término empleado para este programa que realiza la acción de navegar por las diferentes Webs recopilando la información de cada una de ellas tiene varias acepciones, **Crawler** es el más utilizado a nivel general, aunque también podemos hacerle referencia como Spider, o en español como Zombi o Araña.

El Crawler por lo general, es un programa sencillo que funciona lanzando una petición HTTP a la URL semilla, guarda el contenido de la respuesta y analiza todos los enlaces a páginas o contenido externo y realiza el mismo procedimiento con los enlaces extraídos. Esto resulta en un árbol estructural donde cada nodo es un enlace, siendo el raíz la **URL semilla** y todos los demás por debajo de él Webs que podemos llegar partiendo de esa misma. Este árbol obtenido puede ser inmenso e incluso haya información o enlaces que no sean interesantes, por ello todos los Crawlers suelen tener una serie de reglas que concretan qué enlaces son de interés y cuáles no, reduciendo en gran medida la anchura y profundidad del árbol.

2.2.2 Tipos de implementación

Dependiendo del modo en que se recorra el árbol de Webs, se puede implementar el Crawling de dos maneras:

- **Recorrido en Anchura**
- **Recorrido en Profundidad**

Estas estrategias son habituales en el mundo del Crawling. La primera se basa en recorrer el árbol en anchura, visitando primero las Webs a las que hace referencia la semilla y después inspeccionar todos los hijos de cada una de estas. En resumen, no se visitará ningún nodo del nivel N hasta haber visitado todos los nodos del nivel $N-1$. Por el contrario, los Crawlers con recorrido en profundidad son aquellos que, para un nodo, examina a dónde le lleva cada hijo hasta que no haya más hijos (ha alcanzado un nodo hoja), y no pasa al siguiente hasta que no termine con el anterior.
Estos dos tipos son los primeros que se idearon, aunque ninguno de los dos es el óptimo. En la actualidad hay varios estudios que proponen, por ejemplo, no inspeccionar el árbol de referencias entero, ya que los enlaces que se sitúan en profundidades muy bajas pueden no aportar información sobre los contenidos en el nodo semilla. También hay varias publicaciones que proponen métodos en cuanto a selección de los nodos que debemos visitar[16][17].

2.3 Algoritmos de Similitud

2.3.1 Definición

Los algoritmos de similitud son aquellos que, a partir de efectuar cálculos matemáticos, son capaces de evaluar el grado de semejanza entre dos objetos. En matemáticas se pueden utilizar algunos de estos algoritmos para expresar, por ejemplo, la distancia entre dos puntos en un plano, distancia entre dos vectores, etc. Como en este proyecto, y en general en el ámbito del Procesamiento del Lenguaje Natural[18], es interesante conocer como de parecidos son dos textos, estos algoritmos se usan para conocer el grado de coincidencia entre ellos.
Hay una gran variedad de algoritmos de similitud, algunos brindan el grado de semejanza entre dos palabras, como el algoritmo de La Distancia de Hamming, que propone que la distancia entre dos palabras es el número mínimo de ediciones requeridas para pasar de la palabra *A*, a la palabra *B*. Aunque haya varios, en este proyecto nos fijaremos en aquellos que puedan representar la similitud entre dos textos completos. En el siguiente apartado se definen los que se implementarán en este proyecto.

2.3.2 Medidas de similitud habituales en texto

2.3.2.1 Similitud Jaccard

De todos los algoritmos seleccionados para el cálculo de similitudes, Jaccard es el más sencillo de entender. Este algoritmo premia a los textos que tengan más palabras coincidentes con el texto principal a examinar. La forma de calcular este índice se rige por la siguiente ecuación:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Esta medida dará puntuaciones comprendidas entre 0 y 1, donde 0 será que la coincidencia entre textos no existe, y 1 cuando los dos textos que estamos valorando son idénticos.
Con respecto a los valores de A y B, estos serán el conjunto de palabras que contiene el texto A y el conjunto de palabras que contiene B, respectivamente.

2.3.2.2 Similitud coseno entre dos vectores

Otra forma de calcular la similitud entre dos textos pasa por su previa transformación a vectores, ya sean de características o de documento los cuales explicaremos en los siguientes apartados, y medir la distancia a la que están separados dichos vectores.

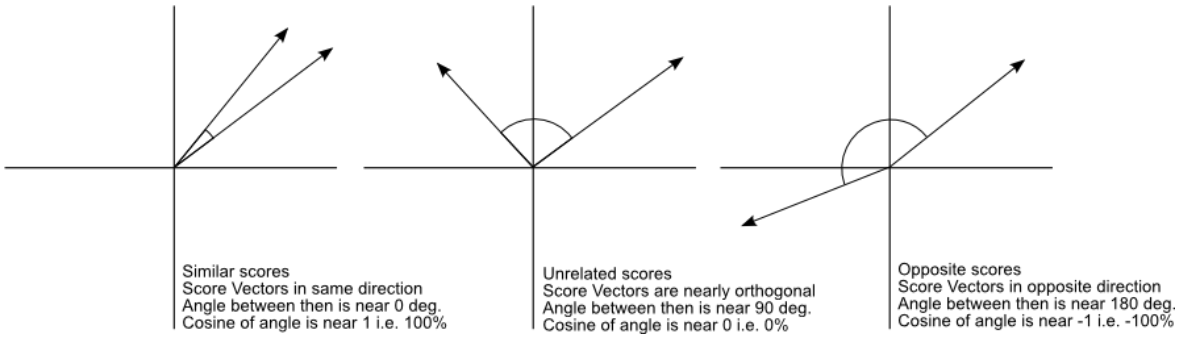

Figura 2-3: Representación de diferentes distancias coseno. Extraído de [19].

De esta forma, se puede apreciar que cuanto más cerca esté uno del otro, más coincidencia tendrán ya que la distancia entre ellos es menor, y cuanto más alejados estén menor será la similitud entre ellos.
La ecuación que nos permite cuantificar la distancia entre dos vectores de iguales dimensiones es la siguiente:

$$sim\_\cos(a,b) = \frac{\sum_{i,j}^{N} a_i b_j}{\sqrt{\sum_i^N a_i^2}\sqrt{\sum_j^N b_j^2}}$$

Los resultados de esta ecuación están comprendidos entre 0 y 1, los cuales representarán la distancia que separa estos vectores.
Hablando de los valores a y b, estos serán los vectores de características o vectores de documento de un texto.

2.3.3 Representaciones vectoriales de textos

Este apartado explica qué métodos se han utilizado para transformar un texto a un vector para así poder hacer cálculos entre ellos. Esto es muy importante entenderlo ya que da la idea principal de cómo se puede calcular la similitud entre dos textos utilizando un elemento tan matemático como los vectores.

2.3.3.1 Cálculo de vectores de características por Word Embeddings

Los *Word Embeddings*[20] son una técnica bastante novedosa que pretende generar un vector de $\boldsymbol{n}$ dimensiones a partir de cualquier palabra. Esta técnica del *Procesamiento del Lenguaje Natural* utiliza el contexto en el que reside nuestra palabra para determinar la representación vectorial de ella. El tipo de *Word Embedding* que más se utiliza debido a su alta probabilidad de acierto es el **Word2Vec**. La forma en que este modelo utiliza el contexto es en base a **ventanas de contexto**, que son la cantidad de palabras que tendrá nuestro contexto más la **palabra de enfoque**, que es la palabra de estudio. Por ejemplo, si nuestra ventana es de 6 palabras, nuestro contexto estará formado por la palabra de enfoque, las 3 anteriores y las 3 posteriores a esta.

El modelo de **Word2Vec**, creado por *Tomas Mikolov* en el año 2013, utiliza una **red neuronal poco profunda**, con dos capas, para obtener el vector de una palabra el cual viene dado por el valor de los pesos que la red neuronal tenga una vez entrenada.
En el artículo *"Efficient Estimation of Word Representations in Vector Space"*[21] de *Mikolov, T.* junto a otros investigadores realizaron un estudio para desarrollar un modelo el cual maximizara la similitud entre vectores de palabras. En este artículo se prueban varias redes neuronales como **redes neuronales con retro-propagación**, **redes neuronales recurrentes** y **redes neuronales simples poco profundas**. Después de realizar varios experimentos, llegaron a la conclusión de que, para llegar a una tasa de acierto alta, es necesario poder entrenar con un gran dataset la red neuronal, por ello, los modelos de redes neuronales que funcionan mejor son los **simples**. En esta publicación, *Tomas Mikolov* propone dos modelos de arquitectura para el aprendizaje de la red. Estos son ***Continuous Bag-of-Words (CBOW)*** y ***Continouous Skip-gram***.

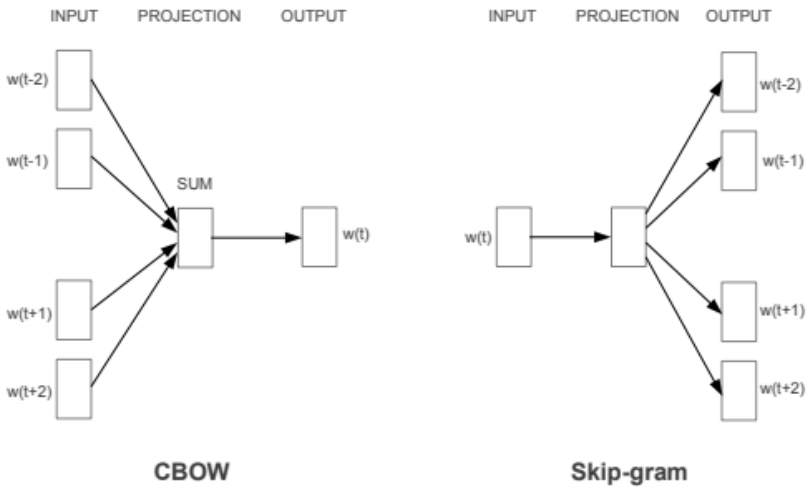

Figura 2-4: Modelos de aprendizaje CBOW y Skip-gram. Extraído de [21].

El modelo de ***CBOW*** es el modelo más intuitivo, donde dada una ventana de contexto, se tiene que adivinar la palabra de enfoque. El modelo ***Skip-gram*** sin embargo, intenta predecir las demás palabras del contexto a partir solo de la palabra de enfoque. De primeras, el modelo ***CBOW*** al ser más intuitivo parece mejor, pero como la red neuronal va a entrenar con una colección de datos muy grande, se puede obtener mucha más información utilizando el modelo de aprendizaje ***Skip-gram***.
En el artículo *"Distributed Representations of Words and Phrases and their Compositionality"*[22], también de *Tomas Mikolov*, se hace una mayor explicación sobre el modelo ***Skip-gram***, la matemática que hay detrás de este modelo y más resultados obtenidos por él.
Otros ejemplos de *Word Embeddings* diferentes al **Word2Vec** desarrollado por **Google** son **FastText**[23], librería de código abierto desarrollada por **Facebook** en **2016** y **GloVe**[24] (***Global Vectors for Word Representation***) desarrollado en la **Universidad de Stanford** en **2014**.

2.3.3.2 Cálculo de los vectores de documento por tf-idf

La idea base del método de vectorización de textos *tf-idf*, con siglas en inglés *Term frecuency – Invese document frecuency*, propone crear vectores de documento los cuales expresen la relevancia de las palabras para cada documento de una colección.

Un vector de documento calculado por *tf-idf* tendría la siguiente forma:

$$\bar{d} = \langle tfidf_1, tfidf_2, tfidf_3, \dots, tfidf_n \rangle$$

Estos vectores tendrán tantas dimensiones como palabras distintas haya en la colección de documentos y donde cada $tfidf_i$ se calcula de la siguiente manera:

$$tfidf_i = tf_i * idf_i$$

Empezando por el primero de los coeficientes, tf marca la frecuencia de un término dentro de un documento y se calcula de la siguiente manera:

$$tf(t,d) = 1 + \log_2\big(frec(t,d)\big)$$

Donde $frec(t,d)$ es la frecuencia del término t en el documento d.
Por otra parte, el segundo coeficiente que compone una dimensión del vector de documento, idf define la frecuencia inversa de documento, lo que propone calcular si el termino elegido es relevante dentro de la colección de documentos. El cálculo de este valor se define como el cociente entre el número total de documentos en la colección entre el número de documentos en los cuales aparece el término estudiado:

$$idf(t,D) = \log_2\left(\frac{|D|}{frec(t,D)}\right)$$

Una vez vista la teoría de esta forma de vectorización concluimos el apartado haciendo una pequeña reflexión sobre lo que supone el valor *tf-idf*.
Para un término t y un documento d dados, **el valor de $\boldsymbol{tfidf_t}$ puede ser menor** en los casos donde el término sea muy utilizado en el documento, pero no en el resto de la colección, ya que el valor de tf sería un valor muy grande, pero idf tan pequeño que penalizaría al anterior. En los casos contrarios, donde el término sea poco relevante en el documento dado, pero sí que aparezca en muchos documentos de la colección, en este caso tf tendría un valor bajo e idf alto. **Los altos valores de $\boldsymbol{tfidf}$** se darán cuando tanto la relevancia del término dentro del documento, como la aparición del término dentro de cada documento de la colección, sean altos.

2.3.3.3 Cálculo de los vectores de documento por Bag of Words

La idea de este cálculo es expresar los textos como la ocurrencia de palabras que contiene el mismo. El termino *bag*, bolsa en español, hace referencia a que en este modelo de representación no interesa la estructura del documento en sí, sino las palabras que contenga. Hablando un poco más en profundidad, la vectorización de un documento por el método *Bag of Words* es crear vectores de documentos donde cada dimensión exprese la frecuencia de cada término dentro del mismo documento. Los vectores creados por el método *Bag of Words* tienen la siguiente forma:

$$\bar{d} = \langle BoW_1, BoW_2, BoW_3, \dots, BoW_n \rangle$$

El vector $\bar{d}$ tendrá tantas dimensiones como términos haya en la colección.
La forma en la que se calcula el valor BoW_t será la siguiente:

$$BoW(t,d) = \begin{cases} frec(t,d)\ si\ t \exists d \\ 0\ si\ t \nexists d \end{cases}$$

Expresando una idea final sobre este tipo de vectorización, cada texto de la colección tendría sus propios focos donde se concentran las palabras clave del texto.

2.4 Herramientas relacionadas

En este apartado se hablará de las herramientas que, aunque no realicen exactamente la misma función que la que se desarrollará en este proyecto, sí que realizan tareas similares en cuanto a la búsqueda y obtención de objetos similares a uno o a un grupo de ellos dentro de una colección de objetos. Algunas de estas herramientas son los **gestores de contenido** y los **motores de búsqueda**.

Empezando por los primeros, un **gestor de contenido** es un sistema informático el cual cuenta con una base de datos que brinda información a los usuarios y propone que el desarrollo del diseño y el contenido sean independientes[25]. Esto permite conservar una misma disposición de la información en toda la Web y crea una experiencia de usabilidad positiva para el usuario. El gestor de contenido más extendido y utilizado por muchas Webs de información, como ***The New York Times***, ***Reuters*** y ***CNN***[26], es **WordPress**[27].

WordPress es una plataforma de código abierto que ayuda a los usuarios a tener una Web visualmente buena y de fácil navegación, ofreciendo temas, plantillas, plugins y otros tipos de elementos para el desarrollo de esta. Por otra parte, y la que más interesa para este proyecto, es cómo los gestores de contenido guían a los editores de la página para añadir más valor a la Web por medio de contenidos relacionados a los ya propuestos. Esto lo hacen posible los algoritmos de recomendación, los cuales, a partir de unas etiquetas o temáticas definidas en la Web, buscan y ofrecen información o contenidos relacionados para así incrementar el valor de la página.

Por otro lado, los **motores de búsqueda**[28], como **Google**, **Bing** y **Yahoo!** entre otros, buscan Webs que puedan resolver una consulta dada por un usuario. El funcionamiento de la mayoría de ellos es mediante la búsqueda de la consulta en el índice del buscador. El índice recoge muchas Webs cada una apuntando a información sobre su contenido, de esta forma se puede comparar la consulta del usuario con los elementos de las Webs para dar unos resultados acordes con la petición.

Alejándose del ámbito informático y yendo al territorio humano, podemos identificar a los **Documentalistas** de la redacción de un periódico. Estas personas se encargan de etiquetar las noticias que se publican en base a una o varias temáticas. Su trabajo es muy importante ya que permite a los usuarios investigar sobre los sucesos de un tema en concreto.

3 Análisis y Diseño

En este Trabajo de Fin de Grado se ha desarrollado una aplicación con la finalidad de crear una sucesión de noticias en el tiempo en base a una noticia propuesta. Para llevar a cabo esta labor, se han realizado dos desarrollos en línea, uno para poder obtener noticias de un periódico seleccionado, pudiendo tener así un banco de datos con el que trabajar, y el segundo desarrollo utilizaría los resultados del primero para poder lograr el objetivo de construir un historial de noticias.

3.1 Análisis

Como se han realizado dos desarrollos, se describirán los requisitos funcionales y no funcionales de cada uno de ellos.

3.1.1 Requisitos Funcionales

3.1.1.1 Módulo de recolección de noticias

- RF: La aplicación será capaz de extraer la información de todas las noticias redactadas en un día seleccionado por uno de los periódicos propuestos elegidos, siendo estos: **El País**, **El Mundo**, **20 Minutos**, **El Confidencial** y **Marca**.
- RF: La aplicación será capaz de extraer la información de todas las noticias redactadas en un mes seleccionado por uno de los periódicos propuestos elegidos.
- RF: La aplicación será capaz de extraer la información de todas las noticias redactadas en un rango de fechas seleccionado por uno de los periódicos propuestos elegidos.
- RF: La aplicación hará uso de Spiders para navegar a las distintas páginas de la Web.
- RF: La aplicación será capaz de seleccionar las noticias propias de la web, examinando el DOM, para su posterior revisión.
- RF: La aplicación será capaz de rechazar noticias para su tratamiento si estas necesitan algún tipo de subscripción o método de pago para poder acceder a su contenido.
- RF: La aplicación será capaz de rechazar artículos si estos pertenecen a un subdominio de uno de los dominios de los periódicos propuestos.
- RF: La aplicación hará uso de un scrapper para la recolección de información de una noticia.
- RF: La aplicación usará el DOM de la página de una noticia para extraer la información de ella.
- RF: La aplicación será capaz de extraer los datos relevantes de una noticia en cualquiera de los periódicos propuestos.
- RF: La aplicación guardará los datos recogidos en un fichero JSON.

3.1.1.2 Módulo de construcción del historial de noticias

- RF: La aplicación será capaz de brindar el historial de noticias antecesoras similares a una noticia dada.
- RF: La aplicación será capaz de dar el resultado de la similitud entre los atributos elegidos (máximo dos) de la noticia a analizar y todas las demás.

- RF: La aplicación será capaz de eliminar las palabras o elementos no relevantes en un texto a través de la técnica de omisión de *stopwords* para optimizar el cálculo de similitud entre textos.
- RF: La aplicación será capaz de modificar algunas palabras mediante la técnica de *stemming* para optimizar el cálculo de similitud entre textos.
- RF: La aplicación será capaz de extraer y brindar los textos, tratados o no, de cualquiera de los atributos de una noticia dada.
- RF: La aplicación tendrá la capacidad de descartar las noticias con fecha posterior a la noticia de estudio.
- RF: La aplicación hará uso de herramientas sobre procesamiento del lenguaje natural para que el tratamiento de textos sea óptimo.
- RF: La aplicación podrá calcular la similitud entre el contenido de los atributos de la noticia a estudiar y otra noticia diferente dentro del conjunto de noticias recopilado, mediante los algoritmos de similitud seleccionados. Los algoritmos que se utilizarán en este proyecto serán **Similitud Jaccard** y **Similitud Coseno**, donde con la última se probarán varias formas de vectorización de documentos como **Word Embeddings**, **tf-idf** o **Bag of Words**.
- RF: La aplicación podrá construir un sistema por el cual pueda separar las noticias en rangos de fechas preestablecidos y calcular las similitudes entre las noticias que estén contenidas en estos grupos. Primero se analizarían las más nuevas hasta llegar a las más antiguas, para que la información a estudiar sea más precisa.

3.1.2 Requisitos No Funcionales

3.1.2.1 Módulo de recolección de noticias

- RNF: El módulo contendrá un script que permitirá recolectar toda la información de todas las noticias publicadas por todos los periódicos propuestos escritas en un rango de fecha que proporcionemos.
- RNF: El módulo contendrá un script que nos permita borrar todos los ficheros con datos de noticias que hayamos creado hasta el momento.

3.1.2.2 Módulo de construcción del historial de noticias

- RNF: Tras conocer las noticias similares a una dada, poder construir la historia de la misma atendiendo a la variable temporal de las publicaciones.

3.2 Diseño

3.2.1 Módulo de recolección de noticias

En este desarrollo, como se ha comentado en apartados anteriores, el objetivo principal será poder obtener un cúmulo de noticias, cada una representada por una serie de atributos esenciales que las definen. Para hacer esto posible, a continuación se explicará la estructura que tendrá este módulo y el por qué se han tomado esta y otras decisiones.

3.2.1.1 Estructura general

La siguiente imagen describe la estructura desarrollada.

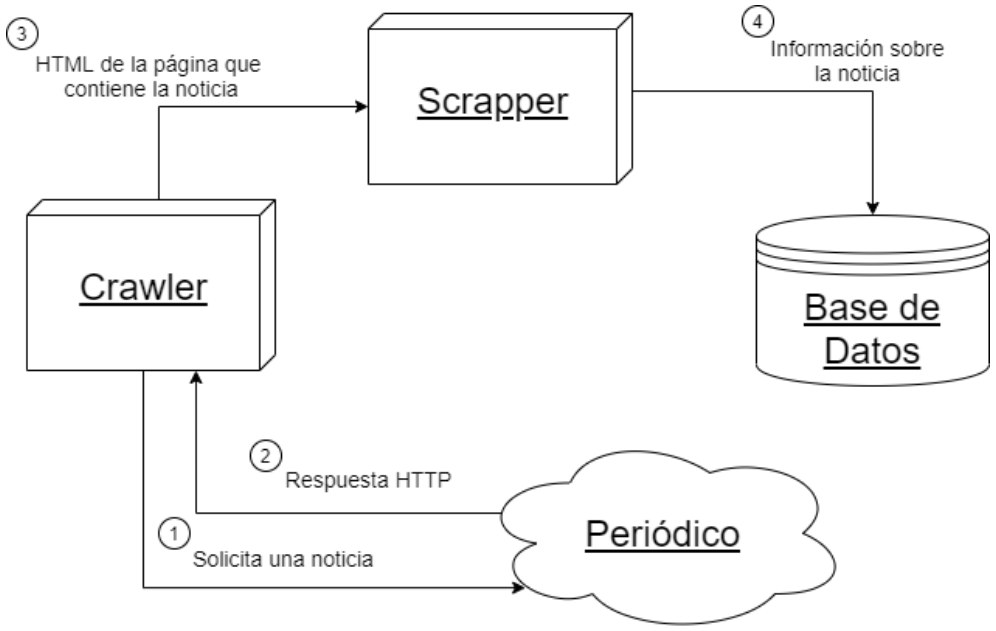

Figura 3-1: Esquema de la estructura del módulo de recolección de noticias

Como se ve, dispone de cuatro grandes entidades, tres de ellas definidas por nosotros mismos, el **Crawler**, el **Scrapper** y la **Base de datos**, y una cuarta que será la **Web del periódico** del cual se desean extraer las noticias.
Para la navegación entre páginas de la Web de noticias, se hará uso de **Spiders o Crawlers**. Estos programas serán los que hagan peticiones al servidor del periódico, examinen el DOM de las diferentes portadas de noticias y extraigan los enlaces de las noticias contenidas en ellas.
La forma en la que se examinarán las páginas de noticias obtenidas será mediante un **Scrapper**. Este programa es capaz de examinar la estructura HTML de la página de la noticia y mediante reglas, extraer los atributos que se hayan definido.
Finalmente, habrá una **Base de Datos** donde se guardarán las noticias examinadas o, mejor dicho, los atributos que definen a cada noticia.
Por último, la decisión de llevar a cabo así la estructura de este módulo es porque la implementación de un *crawler* y un *scrapper* es sencilla y eficaz, más aún con la herramienta elegida para desarrollarlos que será comentada en los próximos apartados. También es debido a que es una forma de mantener dos funcionalidades separadas una de la otra, donde la única interacción entre ellas es el envío de información para que sea procesada.

3.2.1.2 Decisiones para la extracción de noticias

Al principio del proyecto se pensó que la mejor manera de obtener noticias era hacer *crawling*, tomando como semilla la página principal del periódico. Esta idea es bastante trivial, ya que, definiendo las reglas necesarias, se podría llegar al objetivo. Aunque mirando los aspectos negativos, esas reglas que hay que definir podrían ser muy complicadas ya que se tendrían que hacer varios saltos de navegación, pudiendo conllevar a varios errores y mucho tiempo de ejecución.
Pensando en este problema, los periódicos deben tener una forma de almacenar todas las noticias de modo que cualquier usuario interesado pueda consultar cualquier suceso en el pasado. Es primordial que un periódico informe a las personas de lo que pasa en la actualidad y de los temas del pasado, por ejemplo, para tareas de documentación de sucesos. Esto es lo que proponen las **hemerotecas**.

La **RAE** define las **hemerotecas** como *"bibliotecas en que principalmente se guardan y sirven al público diarios y otras publicaciones periódicas"*[29]. Los periódicos en sus versiones digitales presentan una hemeroteca donde el usuario puede seleccionar un día en específico (siempre una fecha posterior a la creación de dicha hemeroteca) para informarse de los sucesos que acontecieron ese día. Además, los periódicos suelen dividir los días en ediciones que pueden ser *mañana*, *tarde y noche*. Por tanto, si se selecciona la edición de un día en concreto, la Web del periódico brindará la portada con los sucesos de ese día.
Finalmente, la forma en la que se recolectarán noticias será, para cada día en el rango de fechas seleccionado, coger cada edición y examinar todas las noticias que contenga.

3.2.1.3 Atributos que definirán una noticia

Hay diversidad de opiniones en cuanto a qué atributos definen una noticia, algunos pueden ser más importantes que otros, unos pueden no aportar nada de información, etc. En este proyecto, se ha decidido que los atributos que definan una noticia sean:

- **Titular**.
- **Enlace a la noticia**.
- **Palabras clave o *keywords***.
- **Resumen**.
- **Autor/es**.
- **Localización**.
- **Fecha de publicación**.
- **Pie de foto**.
- **Firma del pie de foto**.
- **Cuerpo**.
- **Tags**.

Todos estos atributos son autodescriptivos, aunque puede haber dudas de la diferencia entre **keywords** y **tags**. El primero suele ser un metadato y representa las palabras que más se repiten y definen mejor la noticia, mientras que el segundo son etiquetas puestas por el **Documentalista** las cuales tratan de encasillar la noticia en unos temas.

3.2.1.4 Decisiones para el almacenamiento de información

En el proyecto se pensó en cuál sería la forma óptima para guardar los datos de las diferentes noticias obtenidas. Primero se vio la idea de montar una base de datos no relacional con **MongoDB**[30], ya que montar una base de datos relacional no sería lo óptimo ya que para este desarrollo solo es necesario tener una entidad *"noticia"*. Esta entidad tendría un atributo para cada una las características de esta, además una base de datos no relacional ofrece rapidez en las consultas cuando se tienen grandes cantidades de datos almacenados.
Aunque montar una base de datos podría ser una buena idea de diseño y eficiencia, ha sido descartada para este TFG. Por tanto, se ha decidido que las noticias se guardarán en ficheros JSON. La estructura de estos ficheros será de una lista con tantos objetos como noticias se hayan registrado durante la ejecución del programa. Estos objetos contendrán los atributos que definen a una noticia. Esta decisión se ha tomado porque para este proyecto, el guardado de noticias en una base de datos o no, no aporta nada al resultado final del proyecto. No obstante, el almacenamiento de noticias en una base de datos se comentará en el apartado de **Trabajo futuro**.

3.2.2 Módulo de construcción del historial de noticias

Este módulo desarrollado tiene como objetivo abordar la meta de este proyecto, crear el historial de sucesiones a una noticia. En este apartado del diseño del módulo, se hablará

sobre la estructura global del programa, diferenciando los componentes involucrados y las ideas que han llevado a realizar el módulo de esta forma.

3.2.2.1 Estructura general

La siguiente imagen define la estructura del módulo que se está examinando.

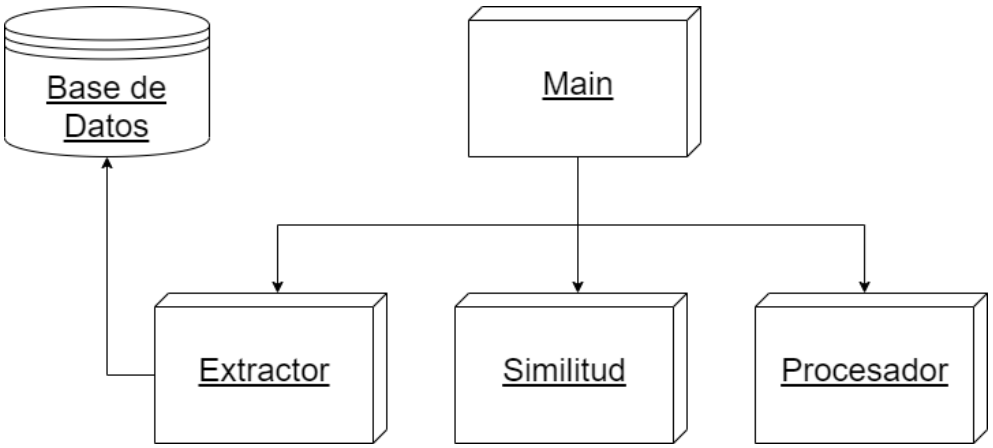

Figura 3-2: Esquema de la estructura del módulo de construcción del historial de noticias

En la imagen se observa que hay cinco grandes entidades que entran en acción en este módulo, una de las cuales conecta directamente con el desarrollo comentado en el apartado anterior, la **Base de Datos de Noticias**.
Empezando por una de estas cuatro entidades desarrolladas en este módulo, está el **Extractor** el cual, como su propio nombre indica, extrae las noticias de la base de datos donde permanecen guardadas las noticias y además trata los textos de estas para su posterior uso. Después se observa la entidad **Similitud** cuya función es calcular la similitud entre los diferentes textos aplicando los algoritmos propuestos en el apartado de **Estado del Arte**. La tercera entidad desarrollada para este módulo será la de **Procesador** donde reside la funcionalidad de guardado de puntuaciones entre una noticia y la noticia de la cual se quiere extraer su historial. Finalmente, la entidad **Main** realizará la función de orquestar todos los componentes anteriores, ya que según se han ido mencionando, el resultado de una es la entrada de la siguiente, y cada entidad se encarga de manejar el flujo principal de datos de la aplicación.
El porqué se ha decidido diseñar de esta manera la aplicación es debido a qué se ha intentado dar un peso importante a la modularidad, que la funcionalidad esté separada de la mejor forma posible. Esto es porque si alguien quiere trabajar sobre este proyecto, la curva de aprendizaje en cuanto al uso de la aplicación se pretende que sea lo más plana posible.

4 Desarrollo

Este apartado es en el que más tiempo se ha invertido en el TFG, debido a los requisitos que se han de cumplir a la hora de construir un historial de noticias.
Como se explicó en el capítulo anterior, se han creado dos módulos, cuyos detalles de implementación se explicarán en este apartado con el enfoque necesario para entender qué herramientas se han utilizado, cómo funcionan los módulos entre sí y cuál es la funcionalidad de cada pieza que compone cada módulo. Asimismo, se explicarán las decisiones que se han tomado y por qué se creen que son las correctas en este proyecto. Después de que se tenga la idea de qué funciones realiza cada módulo y como se ha construido para hacerlas, se expondrá y explicará la ejecución de cada módulo, pudiendo así acabar de entender por completo cada parte.
Para acabar con la introducción del capítulo, al igual que en anterior, primero se empezará enunciando las decisiones tomadas, funcionalidad creada y ejemplos del módulo de recolección de noticias, ya que así se da la visión de cómo se ha trabajado durante el proyecto, ya que el desarrollo empezó por esta parte.

4.1 Módulo de recolección de noticias

4.1.1 Decisiones de Desarrollo

4.1.1.1 Crawling & Scrapping

Sobre este tema se han estudiado varias opciones. La primera de ellas era hacer una implementación con un Crawler, para navegar por las Webs de los diferentes periódicos, y un Scrapper para que una vez extraído los diferentes HTML, se pueda localizar y guardar la información necesaria.
Esta es la manera más extendida de realizar este tipo de tareas, pero investigando se ha dado con una que engloba estas dos funcionalidades. Esto se ha podido realizar ya que se ha utilizado el Framework de **Scrappy**, comentado en el apartado **Estado del Arte** de la memoria. Esta herramienta proporciona varias estructuras, como ficheros de configuración para cambiar parámetros, por ejemplo los encabezados de las peticiones HTTP, tiempos entre peticiones, etc. También proporciona un Middleware, que para este proyecto no se ha usado, pero es una buena herramienta si es necesario realizar alguna funcionalidad antes o después de enviar peticiones a las Webs. También, **Scrapy** cuenta con un constructor de objetos y funcionalidades de escritura en ficheros o en bases de datos. Por último, y el añadido más importante, son las **Spiders**, que funcionan como un Crawler y un Scrapper. Implementan una serie de reglas para restringir accesos a algunas páginas y definir las webs que se quieren visitar, una funcionalidad de **callback**, la cual cada vez que se haga una petición y se obtenga respuesta de una página definida como "interesante", se invocará enviando como parámetro el código HTML de la página en cuestión. Así, teniendo la respuesta HTML de la página, se inspeccionará para extraer los datos que se precisen.
Resumiendo, la idea de coger **Scrapy** como herramienta es por el potencial que ofrece en su funcionalidad, pudiendo desarrollar dos módulos englobándolos en uno solo, además de que el manejo de la herramienta es muy sencillo y no supone dificultades en su uso.

4.1.1.2 Referencia de datos en páginas HTML

Como se verá más adelante en este capítulo de la memoria, el módulo hará peticiones a las páginas Webs de los diferentes periódicos para obtener las noticias como páginas HTML. En el apartado **2.1.2** de la memoria se exponen dos formas de extraer información dentro de

una página HTML, estas son **Selectores CSS** y **XPath**. Después de valorar las dos alternativas, se decidió que trabajar con expresiones de **XPath.** Esto se debe a que se ha visto que es más sencilla la forma de crear este tipo de expresiones y que en el momento en que se tengan que actualizar, que como se explica en el apartado de **Trabajo futuro** es necesario realizar esta tarea, se reduzca en la medida de lo posible la dificultad de esta.

4.1.1.3 Spiders

El debate que se plantea en este apartado es decidir cuántas *Spiders* se han de desarrollar ya que hay dos opciones. La primera de ellas sería realizar **un Spider único para todos los periódicos**. Esta opción es buena si lo que interesa tener es toda la funcionalidad recogida en una sola clase. Con solo este beneficio, no compensa tener todos sus demás contras como por ejemplo:

- Demasiada funcionalidad para una sola clase.
- Funcionalidad no modularizada.
- Dificultad de seguimiento de los datos.
- Mejora del rendimiento de la aplicación.

Por ello, la otra opción sería tener **un Spider para cada periódico**. Esto es mejor en cuanto a eficiencia en código ya que, si no se obtienen los datos requeridos, es más fácil detectar el error porque se sabe de cual Web proviene. Por otro lado, en temas de modularidad es un acierto, teniendo el código limpio sin repetir funcionalidades, separándolas en bloques definidos, lo cual es buena práctica.

4.1.2 Estructuración de la funcionalidad

En este apartado se definen los ficheros desarrollados y cuál es la función que desempeña cada uno.

Como para llevar a cabo este módulo se ha utilizado el Framework de *Scrapy*, se debe seguir una estructura de proyecto impuesta por la herramienta, aunque después esta se modifique con el objetivo de crear la funcionalidad precisada. Por ello, los ficheros involucrados son:

- **settings.py**: **Fichero propio de Scrapy**. En él se pueden realizar múltiples configuraciones que se adaptan a múltiples necesidades. Algunas que se han utilizado son:
    - La posibilidad de no tener que realizar la lectura del archivo *robots.txt*[31]. Este archivo define rutas de la propia página Web que los bots de indexación o Scrappers tienen o no permitido rastrear. Durante el desarrollo, se han encontrado algunas Webs de periódicos que restringen el acceso a los Crawlers a las hemerotecas por medio de este archivo. En este proyecto se ha tratado de respetar siempre el protocolo **robots.txt**, buscando que el proceso de scrapping sea lo menos intrusivo posible y no se provoque perjuicio alguno para el medio de comunicación oportuno.
    - Posibilidad de enviar cabeceras personalizadas. Las peticiones HTTP que **Scrappy** hace a las Webs no contienen información en la cabecera. Durante el desarrollo se ha detectado que algunas Webs detectan el uso de *bots* mediante el uso de estas cabeceras. Por tanto, con el envío personalizado de cabeceras a la hora de realizar peticiones, se puede simular a un usuario en un navegador web, para poder así tener acceso de manera automática a la Web.
    - Posibilidad de imponer un tiempo entre peticiones. Al igual que en el caso anterior, los bots suelen realizar peticiones con muy poco tiempo entre ellas. Por tanto, si una Web detecta esto, puede restringir el acceso a quien

esté haciendo estas peticiones tan frecuentes. Si se define un tiempo fijo entre peticiones, este problema se resuelve.

Hay muchas más configuraciones que nos permite este archivo. Entre ellas, deshabilitar cookies, proponer varios *middlewares* con distintas prioridades de uso al igual que varios *pipelines* si se tuvieran, pero para este proyecto solo se necesitará uno el cual se comentará más abajo.

- **middlewares.py**: **Fichero propio de Scrapy**. Proporciona herramientas para crear funcionalidades en puntos concretos de la ejecución de la fase de **crawling**. En concreto, antes o después de realizar una petición a una Web, antes o después de escribir los datos ya procesados, en saltos de excepciones cuando fallan peticiones, etc. Este fichero puede ser relevante en otro tipo de proyectos, pero para este no se ha visto ningún beneficio claro que ayude tanto como para hacer uso de él.
- **items.py**: **Fichero propio de Scrapy**. Indica qué atributos tendrá el objeto a crear, poniendo el enfoque en una noticia, los atributos que la definen. **Scrapy** permite tener varias clases dentro de este fichero que propongan la misma lógica, pero para este proyecto, con tener uno nos valdrá. Los atributos que guardaremos aquí serán:
  - **titularNoticia**: Titular de la noticia.
  - **linkNoticia**: URL para acceder a la noticia.
  - **keywordsNoticia**: Palabras clave de toda la noticia.
  - **resumenNoticia**: Descripción que aparece después del título.
  - **autorNoticia**: Quien o quienes han creado la noticia.
  - **localizaciónNoticia**: De donde proviene la información.
  - **fechaPublicacionNoticia**: Fecha en la que se ha publicado la noticia.
  - **pieDeFotoNoticia**: Resumen situado debajo de la foto que describe la noticia.
  - **firmaDeFotoNoticia**: Quién ha capturado esa imagen o de donde viene el recurso.
  - **cuerpoNoticia**: El texto que redacta los sucesos acontecidos en la noticia.
  - **tagsNoticia**: Temas en los que el **Documentalista** ha encasillado una noticia.
- **pipelines.py**: **Fichero propio de Scrapy**. Elemento que toma partido al principio y al final de la ejecución del programa y cada vez que se termina de registrar los datos necesarios de una página de una Web. Este fichero realiza las tareas de escritura de objetos en un fichero o en una base de datos.
- **spider_XXX.py**: **Fichero propio de Scrapy**. En este documento se implementa una de las clases de del Framework de **Scrapy**, las *Spiders*. Se indica con los caracteres **XXX** ya que, como se ha comentado en el apartado anterior y ampliamos en este, se ha decidido que cada Web tendrá su propio Spider. Esto se debe a que cada periódico tiene una estructura diferente. Aunque los periódicos comparten formas de estructuración, donde la descripción de la noticia va después del título o el cuerpo va en un único contenedor, cada periódico atribuye **IDs** diferentes a los tags o incluso utiliza tags distintos. Como para la extracción de atributos se utilizan expresiones **XPath**, se ha priorizado que todas estas sean lo más generales posibles, sacando la mayoría de los atributos de los metadatos de las páginas (tags <*meta*> en HTML). Aun así, siempre habrá algún dato que precisemos y que estén situados fuera de los metadatos. Las *Spiders* de *Scrapy* almacenan varios atributos que son interesantes comentar para poder entender el funcionamiento de estas. Los atributos son:

    - **name**: String que da nombre a la *Spider*.
    - **allowed_domains**: Array con los dominios donde permitimos al *Spider* hacer Scrapping. Para cada *Spider* se definirá uno, qué será la Web del periódico a analizar.
    - **Start_urls**: Array que contiene las URLs semilla.
    - **rules**: Objeto **Rule** que propone reglas que deciden qué páginas se deben visitar y cuales no haciendo uso del objeto **LinkExtractor**. Este último, define la expresión **XPath** que lleva a la página objetivo, además de restringir rutas dentro de la Web del periódico que no se requiere analizar. Finalmente, el objeto **Rule** define la llamada de *callback* que se invocará cada vez que el servidor responda con una página que se haya catalogado como objetivo. Las páginas objetivo de este proyecto serán aquellas que contengan la información de una noticia.
- **Periodico.py**: **Fichero desarrollado por nosotros mismos.** En él se generan todas las URL semilla que se van a examinar en la ejecución. Como se ha comentado en el apartado de diseño, se cogerán como URLs semilla todas las páginas de tipo portada, localizadas en la hemeroteca, que se hayan publicado entre los rangos de fechas especificados. Como cada periódico tiene su forma de estructurar la hemeroteca de ediciones, se ha tenido que estudiar como poder crear todas las URLs de la manera más eficiente y global posible, intentando no repetir funcionalidad. Además, esta clase se encarga de dar nombre al fichero JSON que contendrá los atributos de las noticias extraídas.

4.1.3 Diagrama de secuencia

Para concluir el bloque, se realizará una explicación del funcionamiento de una ejecución simple que el usuario puede hacer. Con esto, se pretende entender como cada bloque desarrollado entra en acción a lo largo de la ejecución. Para ello, la explicación se apoyará con un diagrama de secuencia de la ejecución del programa. En este ejemplo, el usuario requiere las noticias comprendidas entre el día ***10/01/2020*** y el día ***17/01/2020*** eligiendo el periódico ***EL PAÍS***.

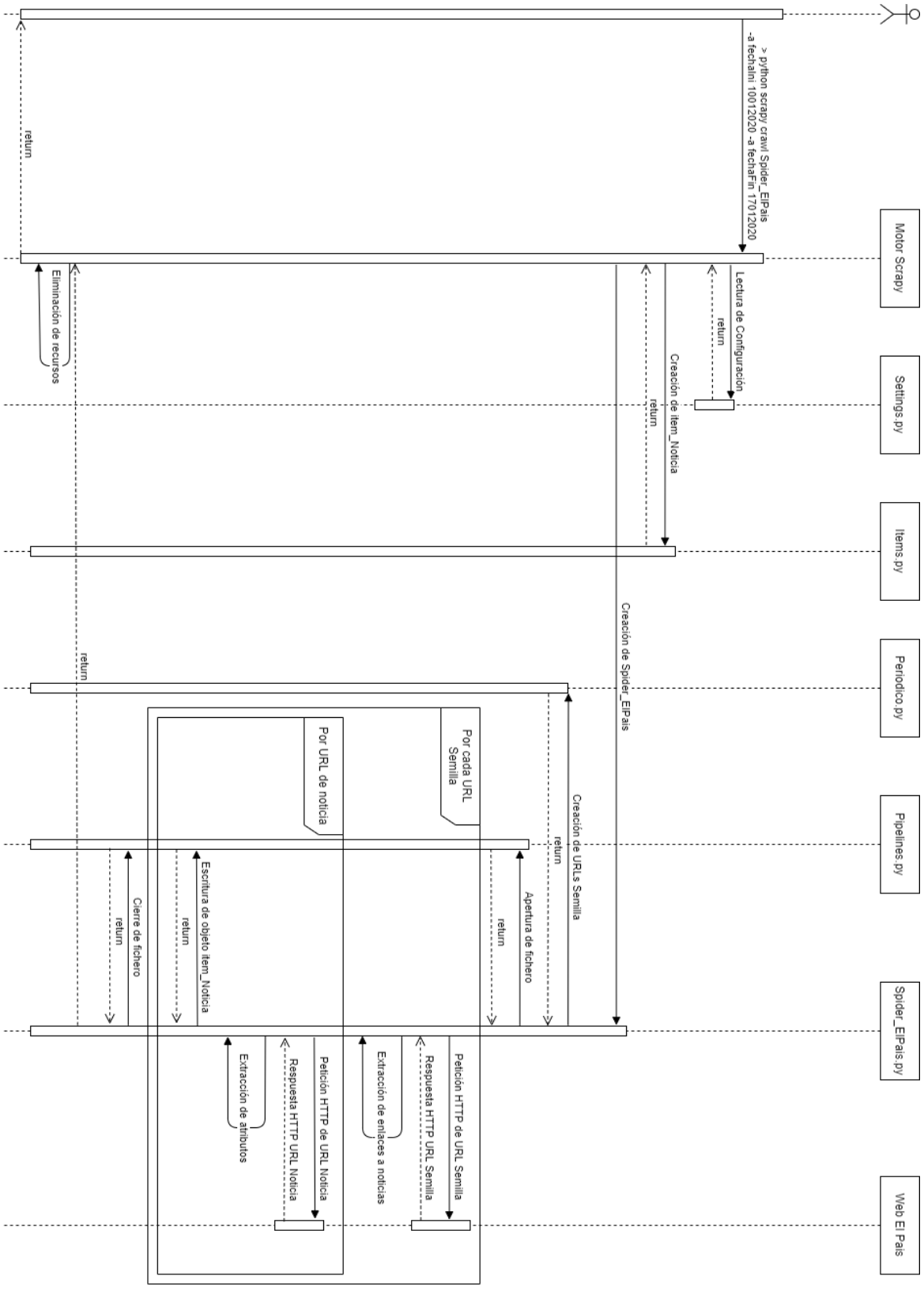

Figura 4-1: Diagrama de secuencia sobre la ejecución del módulo de extracción de noticias

En cuanto el usuario ejecuta el programa, se leen los ficheros de configuración y se crean las definiciones de los objetos creados en *items.py*, en el caso de este proyecto solo uno. A continuación, se instancia el *spider* necesario para extraer noticias del periódico

seleccionado y enviado por parámetros de ejecución, "*spider_ElPais"*. Cuando se construye el *spider* se definen:

- Los dominios permitidos, de los cuales el *crawler* no podrá salir.
- Las reglas de navegación que definen los enlaces interesantes, de los cuales se va a extraer información.
- Páginas dentro de la web a descartar para el análisis.
- Una función de *callback* a la cual se le enviará la respuesta HTTP de una página que se haya definido como importante.

Siguiendo con la ejecución, internamente se comprueba que los parámetros de fecha son correctos y en el caso de serlos, a partir de la creación del objeto **Periódico**, llamando a la función *crea_startUrls()*, se crean todas las **URL semilla** las cuales hacen referencia a cada edición de cada día dentro del rango de fechas proporcionado.

Continuando con la ejecución, a través de la llamada a *super().__init__()*, se invoca al módulo de *pipelines.py* para realizar la apertura del fichero donde escribiremos todos los atributos de las noticias que se van a recoger.

Una vez realizada toda la parte de inicialización, el programa está preparado para lanzar peticiones a las URL semilla, analizar las respuestas y obtener todos los enlaces interesantes, los cuales para nosotros son enlaces a noticias. Como se puede observar, para cada URL semilla se obtendrá el código HTML de dicha página, que se estudiará y se extraerán los enlaces importantes. Si estos enlaces cumplen con las reglas definidas, se enviará una petición HTTP al enlace de la noticia, se estudiará el HTML enviado en la respuesta por el servidor y se escribirán los atributos definidos en *ítems.py* en el archivo definido a partir de la funcionalidad de *pipelines.py*.

Una vez se acabe de estudiar todas las URL semilla, se procederá al cierre del fichero JSON donde hemos escrito todos los datos extraídos y a eliminar los recursos utilizados por el programa aparte de otras tareas internas por parte del motor de *Scrapy*.

Para finalizar este apartado, se debe notificar que en este diagrama no se explica nada sobre el funcionamiento del motor de *Scrapy*, como internamente crea los objetos o manda peticiones al servidor y las recibe. Esto se debe a que un usuario que quiera iniciarse en el uso de este Framework no en necesario que tenga conocimiento de ello. Aun así, en la siguiente figura, extraída de la Web de *Scrapy*, se explica cómo el motor orquesta todos los elementos que entran en juego en la ejecución.

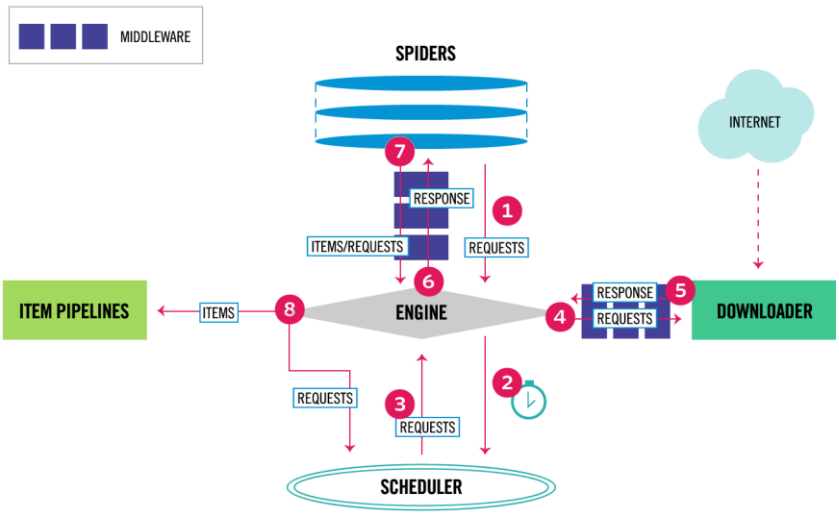

Figura 4-2: Diagrama sobre el funcionamiento del núcleo de Scrapy. Extraído de [32].

4.2 Módulo de construcción del historial de noticias

4.2.1 Decisiones de Desarrollo

4.2.1.1 Tratamiento de textos

El tratamiento de textos es una tarea importante para el desarrollo del lenguaje natural ya que permite la simplificación de estos. Se utiliza en problemas como la creación de índices para mejorar las puntuaciones de similitud entre una búsqueda y un documento. Para realizar un tratamiento existen varias opciones a la hora de modificar palabras o quitar elementos del texto. Algunos ejemplos serían pasar a infinitivo todas las conjugaciones de los verbos, cambiar palabras al género y número más utilizado. Estos ejemplos son en base a la lengua española, que es con la que están escritas la totalidad de noticias que se pueden recoger por los métodos desarrollados en este proyecto.
Para realizar el procesamiento de los textos se pueden seleccionar muchas técnicas, aunque las elegidas para este proyecto y el porqué de estas son:

- **Omisión de signos de puntuación**: Tarea necesaria para la limpieza del texto.
- **Eliminación de *stopwords***[33]: Término anglosajón el cuál hace referencia a las palabras que están en todos los textos y no aportan información. Estas palabras son determinantes, verbos auxiliares…
- **Lematización de palabras**[33]: En este punto existen dos opciones, utilizar *Stemming* o *Lemmatization* para la simplificación de textos. Aunque los dos son buenas opciones a la hora de generalizar palabras y simplificar textos para su posterior análisis, la diferencia entre los dos es que la técnica de *Stemming* simplemente acorta las palabras para quedar con la raíz de estas y la de *Lemmatization* trabaja un poco más las palabras, mirando su análisis morfológico, obteniendo así el *lema* de la palabra o lo que es lo mismo, la forma de la palabra que es aceptada como la forma más simple de todas las que puede derivar. Él punto negativo de la Lematización es el coste en tiempo de ejecución debido a que es una operación más complicada que el *Stemming.* Aun teniendo en cuenta esta última desventaja, hemos decidido **lematizar** los textos para así generalizar textos perdiendo la menor cantidad de información posible.

Por último, se ha decidido hacer uso de la librería **Spacy**[34] de Python para el manejo de textos. **Spacy** ofrece varias utilidades para el ámbito de procesamiento del lenguaje natural. En el desarrollo del módulo se hará uso de el para poder dividir el texto en *tokens* o en palabras y así poder llevar a cabo de forma más eficiente las labores de tratamiento de textos comentadas en este mismo apartado. Además, **Spacy** permite vectorizar textos en base a sus características.

4.2.1.2 Algoritmos de similitud desarrollados

Para este proyecto se ha decidido implementar varios tipos de algoritmos de similitud entre textos. Los algoritmos son los siguientes:

- **Similitud Jaccard:** Algoritmo que toma en cuenta qué palabras coinciden entre textos.
- **Similitud Coseno:** Algoritmo que, dependiendo de la vectorización utilizada, puede puntuar en mayor medida algunos factores. Las vectorizaciones implementadas para el uso de este algoritmo son:
    - **Vectores de características Word2Vec:** Representación de un texto en base a su estructura y a las propiedades de las palabras que contiene. Se utilizará la representación dada por la librería **Spacy**.

- **Vectores de documento tf-idf:** Representación de un texto en base a las palabras que contiene y a su importancia dentro de la colección de documentos.
- **Vectores de documento Bag of Words:** Representación de un texto en base a las palabras que contiene.

Las funciones matemáticas que se utilizarán para el cálculo de similitudes serán las que se explican en el apartado **2.3** de la memoria a excepción de la del cálculo de **idf**, qué para que los datos estén normalizados se usará la siguiente ecuación:

$$idf(t, D) = \frac{|D| + 1}{frec(t, D) + 0.5}$$

4.2.1.3 Metodología para el análisis entre textos

En este apartado se explicará cómo se ha decidido realizar la implementación de comparación de los textos de la colección con el texto de la noticia que estamos estudiando. Primero, antes de mencionar los tipos, los textos que se van a analizar serán los contenidos en los atributos de la noticia. Para hacer la similitud entre textos, **siempre se analizarán entre sí los que provengan del mismo atributo,** además podremos hacer similitud entre textos **combinando hasta dos atributos**, haciendo la concatenación entre los textos.
Hay varias posibilidades de hacer esta tarea, pero nosotros hemos desarrollado dos:

- **Estudio simple**: El **estudio trivial**. Aquí se ha propuesto comparar la noticia seleccionada con todas las demás noticias, una por una.
- **Estudio por fecha**: Este estudio propone **organizar todas las noticias de la colección por fecha en rangos de cuatro horas**, para después **reforzar la noticia a analizar** con más información de las noticias más recientes. La forma en la que se realizará la ejecución será:
    1. División de noticias por grupos.
    2. Computar la similitud entre la noticia seleccionada y las demás noticias pertenecientes a su grupo.
    3. Extraer el texto de la noticia con mejor puntuación y concatenarlo al texto de la noticia a analizar.
    4. Repetir los pasos 2. y 3. con los grupos restantes en orden cronológico inverso.

Este estudio **solo se realizará para los algoritmos de similitud Jaccard y similitud coseno para vectores de características de documento**.
Como conclusión a este apartado, se indica que no se realizará nunca la comparación sobre las publicaciones posteriores a la fecha de la noticia que se está estudiando, ya que el objetivo es encontrar las noticias anteriores a ella.

4.2.2 Diagrama de Clases

En este apartado se mostrará un diagrama de clases para explicar cómo funciona cada componente de este módulo.

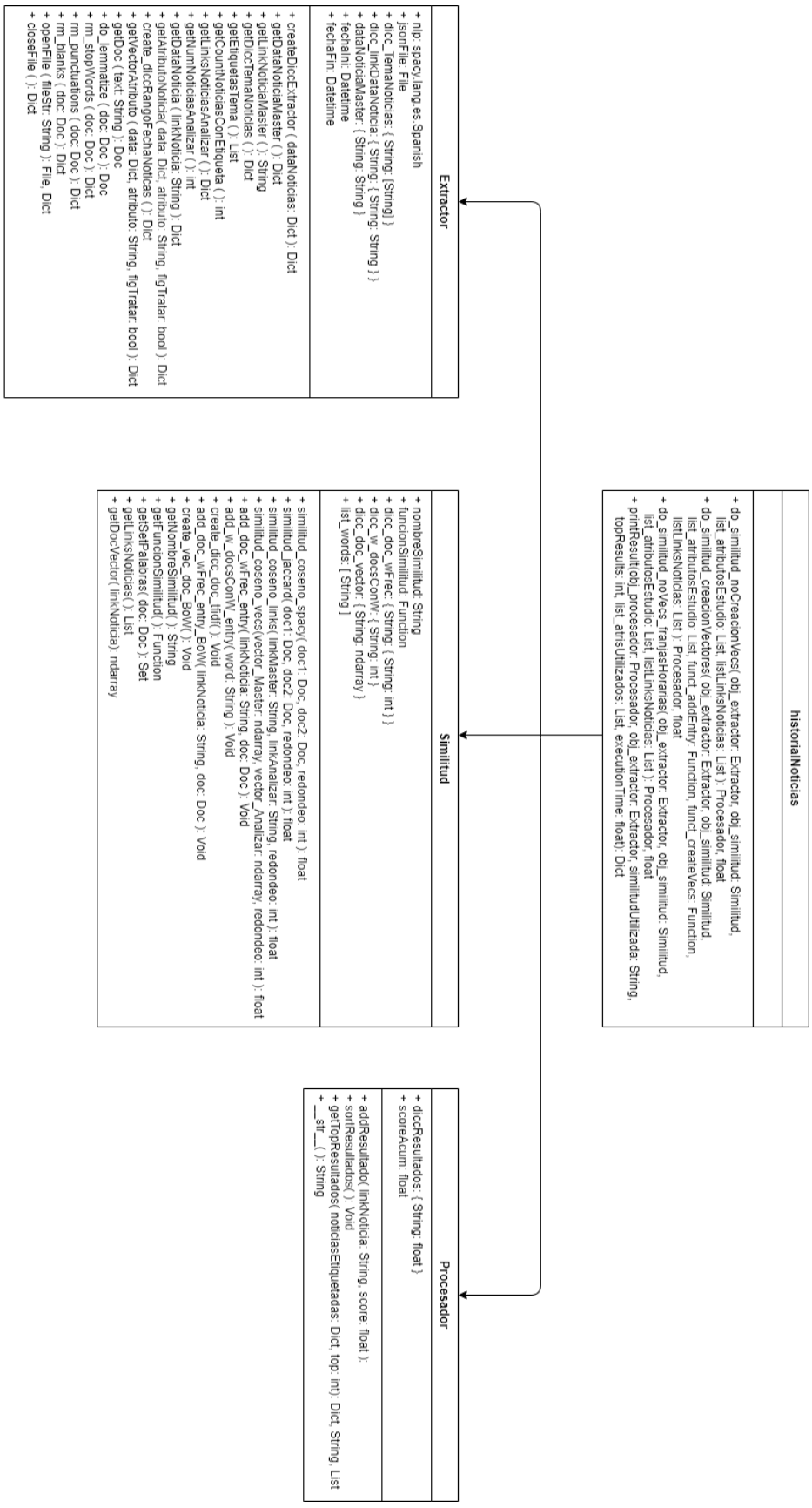

Figura 4-3: Diagrama de clases del módulo de construcción del historial de noticias

Como se puede se observa en el diagrama, hay tres grandes clases que reparten la funcionalidad de todo el módulo en objetivos sencillos y bien divididos, además de un clase que necesita de las otras tres para orquestar la ejecución del programa.
Citando cada clase con una breve descripción del funcionamiento que realiza, tenemos:

- **Extractor**: **Clase que implementamos en el archivo *extractor.py***. En ella reside la funcionalidad de extracción y tratamiento de textos. Este módulo lee los ficheros de noticias resultantes del módulo anterior y extrae la información contenida en ellos para lograr el objetivo de lematizar y quitar los *stopwords* de los textos de las noticias.
- **Similitud**: **Clase que implementa el fichero *similitud.py***. Esta clase implementa todos los algoritmos de similitud comentados en el apartado **4.2.1.2** de la memoria y toda la funcionalidad necesaria para poder calcular los vectores de documento, gestión de vectores y cálculo de conjuntos**.** Esta clase utiliza la librería **Numpy**[35] para el manejo y cálculo de vectores, ya que es una librería estándar y es la más optimizada. Además, para el cálculo de la similitud coseno se ha utilizado la librería **Sklearn**[36], otro estándar.
- **Procesador**: **Clase que implementa el fichero *procesador.py***. Realiza la funcionalidad de guardado de las puntuaciones de la similitud de las noticias de la colección con la noticia elegida para sacar su historial, además de la representación de los resultados obtenidos.
- **Main**: Funcionalidad encargada de orquestar las anteriormente comentadas para obtener el objetivo de este proyecto.

Todas las funciones junto a una descripción de esta, separadas por clase, se puede encontrar en el **Anexo A**.

4.2.3 Pseudocódigo de la ejecución

En esta sección se muestra cómo se ejecuta la aplicación que se ha desarrollado en este apartado para conseguir alcanzar la meta de este proyecto. Para ello, se explicará por fases, primero el caso general que comparten todas las formas de cálculo de similitudes y después cada similitud por separado.

4.2.3.1 Flujo general

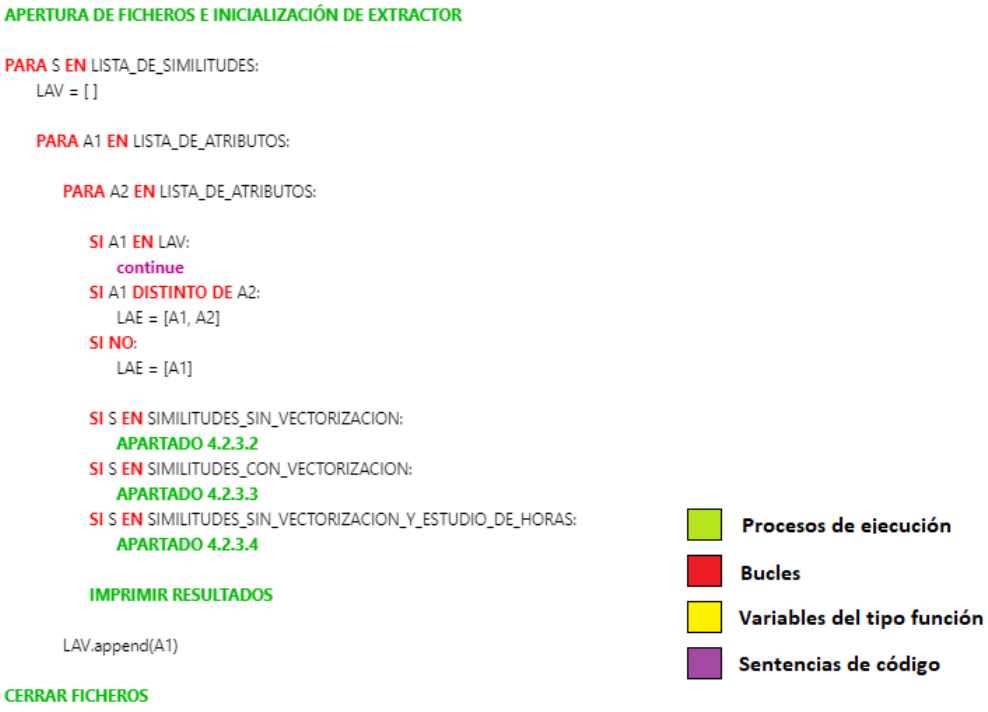

Figura 4-4: Pseudocódigo del flujo general del módulo de construcción del historial de noticias

El primer paso que dé la ejecución del programa será, aparte de abrir los documentos donde se escriben los resultados, crear el objeto **Extractor**, el cual se reutilizará para el cálculo de todas las similitudes. Después se observa un bucle que repetirá la funcionalidad dentro de él 6 veces, siendo estos los algoritmos de similitud implementados. Aquí se inicializa una lista que permitirá conocer si se está repitiendo una combinación de atributos, ya que, para los textos de dos atributos A y B diferentes, se cuenta como el mismo texto la combinación de los atributos de la forma $A + B$ que $B + A$. Siguiendo con el esquema, encontramos dos nuevos bucles, uno anidado al otro, que recorrerán todas las posibles combinaciones de atributos. En este punto hay dos posibilidades:

1. **El atributo del segundo bucle esté presente en la lista de atributos estudiados:** Como se ha comentado antes, se omite esta iteración del segundo bloque para no computar dos veces los mismos resultados.
2. **El atributo del segundo bucle es distinto al del primero:** Como es distinto, se crea una lista con los dos atributos para que, al realizar las similitudes, se concatenen los textos. Si el primer atributo fuera igual al segundo, la lista solo contendrá ese mismo atributo.

Continuando con el esquema, se debe diferenciar entre los tres grandes grupos de cálculo de similitudes y que se explicará con detalle cada caso en los siguientes apartados.
Concluyendo el bucle anidado, se realizará una impresión de los datos obtenidos en un fichero y finalizando el padre de este mismo bucle, se añade el atributo de este bucle a la lista que comprueba la repetición de atributos.
Finalizando el programa, se cierra el fichero de datos abierto por el objeto **Extractor** y los ficheros donde se escriben las puntuaciones de las noticias y el historial.

4.2.3.2 Flujo para algoritmos de similitud sin creación de vectores con estudio simple

Este caso recoge las formas de cálculo de similitud **Jaccard** y **Similitud Coseno para los Word Embeddings de Spacy.** En estos casos, no se necesita un cálculo de vectores previo

al cómputo de las similitudes ya que, para Jaccard solo es necesario un set de palabras fácil de adquirir en Python y los vectores de características de los documentos los proporciona la librería Spacy. Por tanto, la secuencia de ejecución del programa para este caso es la siguiente:

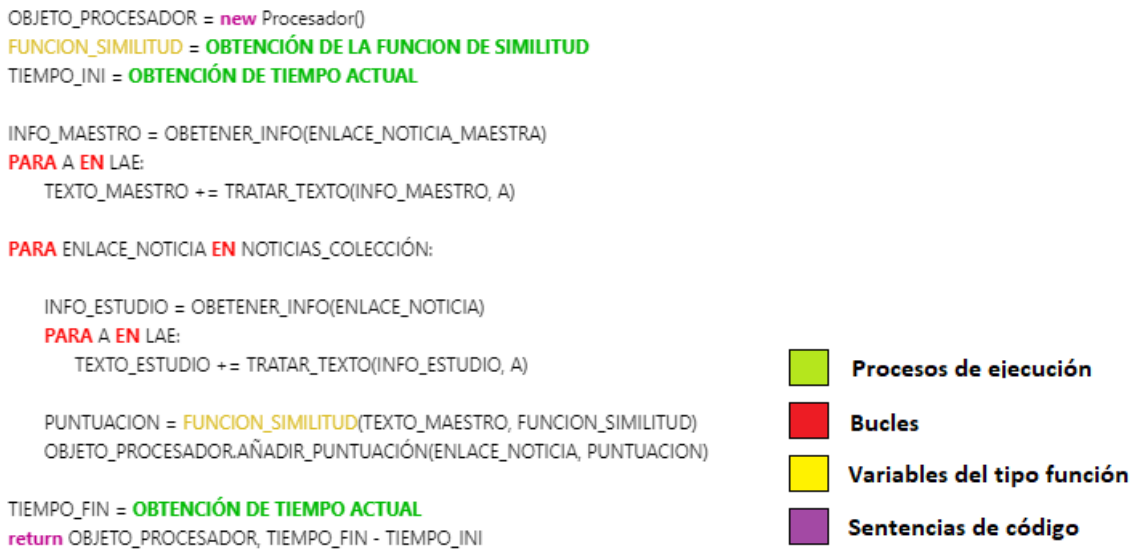

Figura 4-5: Pseudocódigo del flujo para algoritmos de similitud sin creación de vectores con estudio simple

Lo primero que se realiza en la ejecución es crear el objeto **Procesador** para el guardado de resultados y obtener la función de similitud con la que se computan la similitud entre textos. Después, se realiza una llamada a **time** para obtener la hora actual y después conocer el tiempo que tardamos en computar todas las similitudes. El siguiente paso será tratar el texto de la **noticia maestra**, aquí se obtiene el texto correspondiente al atributo seleccionado ya tratado. En caso de que sean dos atributos se concatenan los textos de cada atributo. A continuación, se obtienen todas las noticias de la colección para luego con cada una de ellas, realizar el mismo tratamiento que hemos hecho con la maestra y después calcular la similitud entre los dos textos, guardando los resultados en el objeto **Procesador**. Una vez acabado con este bucle, se calcula el tiempo invertido y se retorna este resultado junto al objeto **Procesador** el cual contiene todos los resultados.

4.2.3.3 Flujo para algoritmos de similitud con creación de vectores con estudio simple

Este caso calcula las similitudes entre textos con el algoritmo de **Similitud Coseno** con las **vectorizaciones tf-idf** y **Bag of Words**. Los vectores de documento que se requieren para aplicar la similitud tienen que ser creados en primera instancia para después poder operar con ellos. Por ello, habrá que recorrer dos veces la colección de documentos. La secuencia de ejecución del programa será la siguiente:

Figura 4-6: Pseudocódigo del flujo para algoritmos de similitud con creación de vectores con estudio simple

Como se puede observar, al igual que en el caso anterior, primero se crea el objeto **Procesador** y se obtiene tanto la función de similitud como la hora actual. El siguiente paso será, para cada noticia, extraer los textos de los atributos seleccionados y añadir los datos necesarios, como palabras contenidas en el texto y su frecuencia o el número de documentos en el que sale cada palabra, al objeto **Similitud.** Una vez acabado este primer bucle, se calculan todos los vectores para cada documento con la función específica de vectorización. Finalmente, se tendrá que recorrer de nuevo la colección de documentos para obtener el resultado entre **todos los vectores de documento** y el vector de la **noticia maestra**. Finalmente, se retorna el tiempo invertido en todo este apartado y el objeto **Procesador** que contiene los resultados obtenidos durante el último bucle.

4.2.3.4 Flujo para algoritmos de similitud sin creación de vectores con estudio por fecha

El último caso que se ha creado para la obtención de resultados es el cálculo de similitud de noticias mediante algoritmos que no requieren vectorización previa y realizando un análisis de noticias por fechas. Como se ha explicado con anterioridad, se realiza este estudio para mejorar los datos de la noticia maestra pudiendo obtener mejores resultados. La secuencia de ejecución se muestra en el siguiente diagrama:

```
OBJETO_PROCESADOR = new Procesador()
FUNCION_SIMILITUD = OBTENCIÓN DE LA FUNCION DE SIMILITUD
TIEMPO_INI = OBTENCIÓN DE TIEMPO ACTUAL

INFO_MAESTRO = OBETENER_INFO(ENLACE_NOTICIA_MAESTRA)
PARA A EN LAE:
    TEXTO_MAESTRO += TRATAR_TEXTO(INFO_MAESTRO, A)

DICCIONARIO_FRANJAS_HORARIAS = CREAR_DICC_FRANJAS_HORARIAS()

PARA LISTA_NOTICIAS_FRANJA_HORARIA EN DICCIONARIO_FRANJAS_HORARIAS.values():

    MEJOR_NOTICIA = ("", 0)
    PARA ENLACE_NOTICIA EN LISTA_NOTICIAS_FRANJA_HORARIA:

        INFO_ESTUDIO = OBETENER_INFO(ENLACE_NOTICIA)
        PARA A EN LAE:
            TEXTO_ESTUDIO += TRATAR_TEXTO(INFO_ESTUDIO, A)

        PUNTUACION = FUNCION_SIMILITUD(TEXTO_MAESTRO, FUNCION_SIMILITUD)
        SI PUNTUACION > MEJOR_NOTICIA[1]:
            MEJOR_NOTICIA = (ENLACE_NOTICIA, PUNTUACION)
        OBJETO_PROCESADOR.AÑADIR_PUNTUACIÓN(ENLACE_NOTICIA, PUNTUACION)

    SI MEJOR_NOTICIA[1] != "" AND SI MEJOR_NOTICIA[1] > UMBRAL_PUNTUACION:
        INFO_ESTUDIO = OBETENER_INFO(ENLACE_NOTICIA)
        PARA A EN LAE:
            TEXTO_ESTUDIO += TRATAR_TEXTO(INFO_ESTUDIO, A)
        TEXTO_MAESTRO = TEXTO_MAESTRO + TEXTO_ESTUDIO

TIEMPO_FIN = OBTENCIÓN DE TIEMPO ACTUAL
return OBJETO_PROCESADOR, TIEMPO_FIN - TIEMPO_INI
```

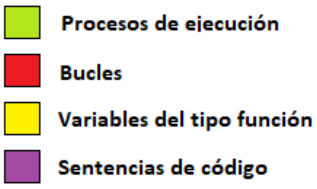

Figura 4-7: Pseudocódigo del flujo para algoritmos de similitud sin creación de vectores con estudio por fecha

Se puede observar que los primeros pasos de la ejecución son iguales que en el apartado ***4.2.3.4*** de la memoria, donde primero se genera un objeto **Procesador**, se obtiene la función de similitud a utilizar, se guarda el tiempo actual y se extrae el texto de los atributos propuestos para el análisis. A continuación, el objeto **Procesador** crea el diccionario de noticias ordenadas por rangos de fecha el cual iteraremos sobre él. Por cada noticia dentro de cada rango obtenido, en orden cronológico inverso, se obtienen los datos de estas y se computa la similitud entre los textos extraídos. Mientras se estudia un rango de fechas, se actualiza una variable que recoge la noticia dentro del rango con mejor puntuación de similitud con el texto maestro. Una vez termina con el rango, se concatena el texto de la noticia con mejor resultado de similitud al texto de la noticia maestra, proporcionándola más información relevante. Una vez se termina con todos los rangos de fecha, se retorna el objeto **Procesador** con los resultados y el tiempo invertido en la ejecución para calcular todas las similitudes.

Para concluir con el apartado de desarrollo, hay que comentar que tanto el **Manual de instalación** para descargar todos los contenidos para ejecutar todos los módulos explicados, como **el Manual del programador** para conocer cómo se ejecuta cada programa desarrollado, se encuentran en los **Anexos B** y **C** respectivamente.

5 Pruebas y resultados

Este capítulo se compone de dos grandes apartados. En el primero se desarrollará la forma en la que se ha valorado la veracidad de la extracción de noticias y los resultados obtenidos después de ejecutar el módulo de extracción de noticias. La segunda parte explica cómo se han definido las pruebas y los resultados obtenidos.

5.1 Módulo de recolección de noticias

5.1.1 Explicación de las pruebas

Para la creación de las expresiones **XPath** y visualización de qué elementos de una Web se escogen, se ha utilizado el *plugin* de **Google Chrome *Xpath Helper***[37]. Esta herramienta permite a los usuarios introducir una expresión **XPath** y ver resaltados los elementos de la Web a los que hace referencia esa expresión. Por ello, esta herramienta se ha utilizado durante el desarrollo para crear las expresiones necesarias para la recolección de todas las noticias de las diferentes portadas de los periódicos, así como para extraer la información de las noticias.

La siguiente figura muestra un ejemplo de las noticias que nuestra *Spider* es capaz de analizar.

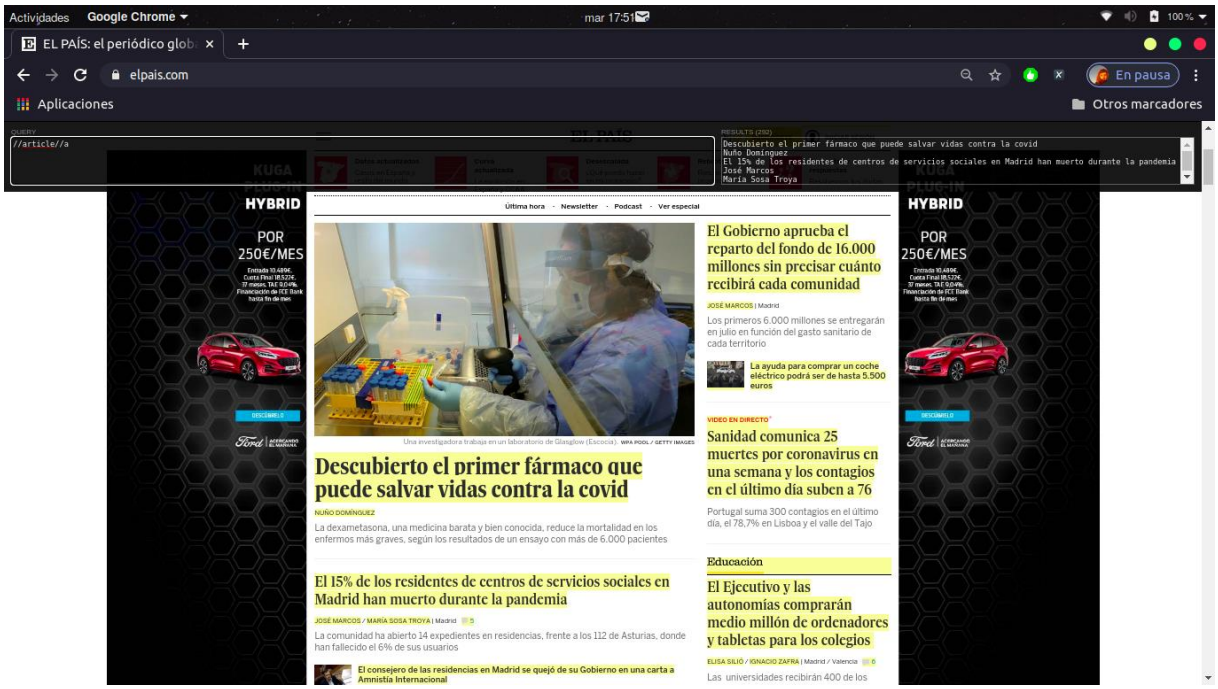

Figura 5-1: Ejemplo de las noticias que se someterán a estudio. Extraído de [3].

Por otro lado, se debe hacer hincapié en que, si la noticia remarcada está en un domino diferente, se prescindirá de ella. Esto se debe a los requisitos funcionales mencionados en el apartado **3.1.1.1.**

5.1.2 Resultados obtenidos

Los resultados obtenidos del proceso de scrapping son los que se muestran en la **Tabla 5-1**:

Fecha	Periódico	Número de noticias extraídas
14/03/2019 – 25/03/2019	El País	1250
	El Mundo	843
	20 Minutos	115
	El Confidencial	1014
	Marca	806
15/04/2019 – 30/04/2019	El País	1709
	El Mundo	1161
	20 Minutos	86
	El Confidencial	1244
	Marca	1001
01/10/2019 – 20/10/2019	El País	1913
	El Mundo	1343
	20 Minutos	93
	El Confidencial	1664
	Marca	1422
10/10/2019 – 27/10/2019	El País	1689
	El Mundo	1100
	20 Minutos	89
	El Confidencial	1493
	Marca	1236
25/11/2019 – 05/12/2019	El País	1085
	El Mundo	601
	20 Minutos	229
	El Confidencial	990
	Marca	707
20/01/2020 – 10/02/2020	El País	2081
	El Mundo	1148
	20 Minutos	126
	El Confidencial	1949
	Marca	1475
05/03/2020 – 20/03/2020	El País	52
	El Mundo	943
	20 Minutos	457
	El Confidencial	1515
	Marca	905

Tabla 5-1: Número de noticias extraídas por fecha y por periódico

Se puede observar que todos los periódicos suelen tener aproximadamente el mismo número de noticias por rango de fecha. Han de hacerse dos apuntes. El primero sería que el periódico **20 Minutos** tiene una cantidad menor de noticias con respecto a los demás periódicos. Esto se debe a que el **20 Minutos**, además de informar sobre temas a nivel nacional, informa también a nivel provincial, las cuales no se quieren en el conjunto extraído. Como este proyecto estudia en un rango de sucesos nacionales o que afecten a todo el país, aunque las noticias se localicen en una zona, de este periódico solo se extraen las noticias de las temáticas **"Economía"**, **"Internacional"** y **"Nacional"**.
El segundo punto se debe a que en los últimos meses, el periódico **El País** está cambiando el dominio de sus páginas dentro del archivo, antiguamente era ***"elpais.com"*** y ahora está

migrando a ***"static.elpais.com"***, por tanto las URLs semilla no se están generando correctamente. Este tema se tratará en la sección de **Trabajo futuro**.
Para concluir con este apartado de pruebas, en la siguiente figura se ha extraído un ejemplo de cómo queda la información de las noticias en los ficheros JSON.


```
{
    "titularNoticia": "Escocia impulsará otro referéndum de independencia para 2021",
    "linkNoticia": "https://elpais.com/internacional/2019/04/24/actualidad/1556113438_335060.html",
    "resumenNoticia": "La primera ministra, Nicola Sturgeon, se decide a dar el salto 72 horas antes (...)",
    "keywordsNoticia": [
        "escocia",
        "impulsar",
        "referéndum",
        "independencia",
        "(...)"
    ],
    "tagsNoticia": [
        "Brexit",
        "Escocia",
        "Referéndum UE",
        "(...)"
    ],
    "pieDeFotoNoticia": "La primera ministra escocesa, Nicola Sturgeon, este miércoles en el (...)",
    "autorNoticia": [
        "Rafa de Miguel"
    ],
    "cuerpoNoticia": "El Brexit, como en tantas otras cosas, es la causa última de que despierte (...)",
    "fechaPublicacionNoticia": "2019-04-24T18:53:10Z",
    "firmaDeFotoNoticia": "Ken Jack (getty) | reuters",
    "localizacionNoticia": [
        "Londres"
    ]
}
```


Figura 5-2: Ejemplo de la estructuración de la información de una noticia una vez analizada

5.2 Módulo de construcción del historial de noticias

5.2.1 Explicación de las pruebas

La forma en la que se ha evaluado el propósito de este proyecto es creando varios **datasets**, uno por periódico, los cuales contienen noticias todas ellas etiquetadas con un tema que las define. Estos son algunos de los temas más representativos del ámbito nacional dentro de los años 2019 y 2020, los cuales se han utilizado para realizar las pruebas:

- **Juicio del Procés**
- **El incendio de Nôtre Dame en París**
- **Las Elecciones Generales del año 2019**
- **La exhumación del dictador Francisco Franco**
- **La Cumbre del Clima del año 2019**
- **La salida del Reino Unido de la Unión Europea, el Brexit**
- **El Día Internacional de la Mujer**
- **La muerte del jugador de Baloncesto Kobe Bryant**
- **La pandemia del COVID-19**

Una vez vistos los diferentes temas con los que se van a trabajar, se discutirán dos puntos que son muy relevantes a la hora de entender el funcionamiento de las pruebas:

- **No se utilizará como atributo válido tagsNoticia.** Esto se debe a que, por temas de eficiencia, se han etiquetado los temas de las noticias en base a si una noticia tiene

un determinado tag en su listado. Esto se ha hecho de esta manera para que el etiquetado sea de una forma automática.

- **No es un problema de clasificación.** Este con diferencia es el punto más importante. Con el fin de validar los resultados obtenidos y que la validación sea automática, se han utilizado herramientas que se usan en problemas de clasificación, aunque el problema con el que nos encontramos en este TFG no es un problema de clasificación. Ya que se está comprobando el grado de similitud entre dos textos, se dirá que las noticias con mejor puntuación de similitud serán las que formen parte del historial de noticias. Este modelo de evaluación no es el ideal, por eso en los resultados obtenidos puede que el algoritmo de similitud puntúe muy alto una noticia que tenga etiquetado un tema diferente al de nuestra noticia, pero luego esta se analice en su totalidad y resulte que trata dos temáticas diferentes. Estas dos serían la asignada en la fase de etiquetado y la tratada en la noticia que se está estudiando.

Una vez definidas estas bases, la función de clasificación que se ha utilizado para medir la calidad de este proceso es la de **Sklearn *classification_report()***[38]. Esta función recibirá dos matrices $(1, N)$, las cuales contendrán la información sobre la temática de las noticias, y un *array* con los nombres de los identificadores de los temas de las noticias, en este caso la temática a estudiar y una etiqueta que hace referencia a todas las demás noticias. De las dos matrices, una será ***y_true***, que contendrá el resultado esperado y otra ***y_pred***, que contendrá el resultado obtenido. Si el identificador de la temática de la noticia que estamos estudiando es el 01 y el identificador para todas las demás noticias es el 00, tendremos:

$$y_{true} = [01, 01, 01, \dots, 01, 00, 00, 00, \dots, 00]$$
$$y_{pred} = [01, 00, 01, \dots, 00, 01, 01, 00, \dots, 00]$$

Como se observa en las matrices, ordenadas por orden decreciente de puntuación de izquierda a derecha, en y_{true} hay 2 secciones. La primera contendrá tantos valores de la temática de la noticia seleccionada como **noticias con el mismo tema se hayan analizado**. Esto se hace así porque se espera que las noticias con mayor puntuación ocupen siempre las primeras posiciones. La segunda sección solo contendrá el valor del identificador de todas las demás noticias y ocupará tantas dimensiones como sea **necesario para alcanzar el total de noticias analizado**. Esto se hace para no evaluar las demás temáticas debido a que este proyecto solo se necesita identificar una temática por ejecución. Por tanto, se observa que las secciones 1 y 2 de y_{true} y la 1 de y_{pred} tienen **el mismo tamaño e igual a la cantidad de noticias del analizadas en la ejecución**.

Por último, en los resultados que nos proporcionará la función de **Sklearn**, se tomará en cuenta el valor resultante de **f1-score**[39], ya que considera los valores de **precision** y **recall** y aporta información sobre los falsos positivos y falsos negativos.

5.2.2 Resultados obtenidos

Debido a la gran cantidad de resultados obtenidos durante el proyecto, en este apartado de resultados, se estudiarán los resultados obtenidos a partir de los datos recopilados del periódico **El País**, evaluando la temática de **La Cumbre del Clima del año 2019 para la noticia que puede encontrarse en la siguiente url: https://elpais.com/sociedad/2019/12/02/actualidad/1575268228_449028.html**. Dentro de este dataset hay **603 noticias**, donde **52** de ellas pertenecen a la temática y anteriores a la de estudio hay **23**.

Analizando los datos resultantes de la **Tabla 5-2**, la combinación **función similitud/ atributos** que ha proporcionado el mejor resultado de **f1-score** es **Similitud Coseno con Vectorización BOW / Keywords**, con **0,7**. Este resultado es bastante bueno ya que, mirando los resultados de similitud obtenidos, vemos que de las 22 noticias ha situado entre las

mejores a **15**. El historial de noticias resultante para el método con mejor puntuación se muestra en la **Tabla 5-3**.

Hablando de cada algoritmo por separado, empezando por la Vectorización con **Word2Vec**, pensamos que no se han obtenido buenos resultados debido a dos razones. Una de ellas el **tratamiento de textos realizado**, ya que esto puede empeorar los resultados debido a que aunque pensemos que estamos mejorando el procesamiento generalizando las palabras a partir de la lematización, le estamos empeorando la capacidad de conocer el contexto ya que estamos modificando las palabras. Por otra parte, ya que la temática escogida puede tener un carácter político, creemos que la segunda razón es debido a que **no es una noticia tan destacada** como otras que hemos estudiado en el proyecto. Además, mirando el estudio por fechas con esta vectorización, vemos que aunque incrementemos la información de los textos, estamos empeorando el contexto por lo que no salen buenos resultados.

La vectorización de textos con **tf-idf** ha funcionado bien, sobre todo cuando los textos a estudiar tienen mucha información. Se puede ver que los mejores resultados obtenidos para este experimento han sido cuando hemos utilizado el **cuerpo de la noticia**, por lo que parece que este algoritmo se comporta mejor con textos abundantes.

La similitud **Jaccard**, además de funcionar bastante bien con el estudio simple cuando los textos contienen el cuerpo, vemos que **el estudio por fecha** aumenta en unos puntos algunos resultados. Esto se debe a que estamos incrementando la información de la noticia, aunque pensamos que con otro método de adicción de información a los textos, podemos mejorar los resultados de todas las combinaciones de atributos.

Por último, pensamos que en este experimento ha salido la combinación de función de similitud / atributos especificada en párrafos anteriores debido a que **BOW** al igual que **Jaccard**, se centra en la similitud entre la información contenida entre dos textos sin importar los demás del dataset, pero en **BOW** se cuenta la frecuencia de las palabras, ya que de la manera en la que se exponen las **Keywords**, hay palabras dentro de una **Keyword** que pueden estar en otra, por lo que ampliamos información.

Antes de pasar con los resultados finales, interesante mirar qué combinación **función similitud / atributos** es la mejor para cada periódico, ya que hay dependiendo de cuál, hay algunos que funcionan mejor con una cierta característica ya que para ese periódico puede tener bastante peso. Las mejores combinaciones por periódico obtenidas, sacadas a partir de los datos resultantes, son:

- **El País: Similitud coseno con Vectorización BOW / Titular + Cuerpo**
- **El Mundo: Similitud coseno con Vectorización BOW / Keywords**
- **20 Minutos: Similitud coseno con Vectorización Word2Vec y Estudio Simple / Keywords + Autor**
- **El Confidencial: Similitud Jaccard y Estudio Simple / Keywords**

Como se ve, cada periódico tiene una distinta combinación de elementos para lograr la mejor puntuación. Esto se ha sacado a partir de todas las simulaciones realizadas, por lo que estos resultados pueden variar a medida que se hagan más pruebas. Además, ya que los resultados no se están obteniendo de la manera ideal, puede que varíen los resultados obtenidos.

Una vez visto un experimento concreto y los resultados por periódico, se comentarán los resultados obtenidos de forma general, los cuales se pueden observar en la **Tabla 5-4.**

Se observa que la combinación de elementos que mejor funciona con cualquier periódico es **Similitud Jaccard y Estudio por Horas / Keywords**. También se han anotado en la tabla la mejor combinación para dos atributos. Además, como el **Estudio por Horas** se ha creado como investigación para ver si se puede conseguir alguna mejora realizando una modificación en alguno de los métodos, se han indicado en verde las mejores puntuaciones obtenidas para los métodos de **Estudio simple** con uno y con dos atributos. Además, como

forma de curiosidad, se han añadido a los lados de la tabla las medias para cada atributo y función de similitud resaltando las que mejor funcionan.
Con estos resultados se observa que, para todos los periódicos, se ha podido crear un historial de noticias con una buena tasa de acierto, por lo que se ha conseguido el objetivo de este proyecto. Aun así, volviendo a hacer hincapié en el segundo de los puntos relevantes comentados en el apartado **5.2.1**, no se pueden tomar los datos como totalmente fiables, debido a que estos pueden variar ya que la evaluación no se ha hecho de la forma ideal. Esto se ha remarcado varias veces durante este apartado debido a que, en algunos historiales, aparecen noticias que han catalogado sobre otras temáticas, pero si se estudia la noticia en cuestión, se haga referencia a esta temática, aunque la noticia tenga otra temática principal.
Los resultados obtenidos en los demás experimentos se pueden observar en el **Anexo D** de este documento.

CUMBRE DEL CLIMA	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,26	0,3	0,3	0,3	0	0,35
Titular + Keywords	0,35	0,35	0,39	0,65	0,22	0,43
Titular + Resumen	0,26	0,39	0,3	0,52	0,22	0,39
Titular + Autor	0,13	0,13	0,3	0,17	0,04	0,17
Titular + Cuerpo	0,13	0,57	0,57	0,57	0,09	0,57
Keywords	0,3	0,39	0,3	0,7	0,22	0,43
Keywords + Resumen	0,26	0,43	0,3	0,52	0,17	0,43
Keywords + Autor	0,35	0,43	0,26	0,61	0,22	0,48
Keywords + Cuerpo	0,22	0,61	0,57	0,52	0,22	0,61
Resumen	0,09	0,17	0,13	0,17	0,09	0,17
Resumen + Autor	0,09	0,22	0,17	0,17	0,09	0,22
Resumen + Cuerpo	0,13	0,57	0,52	0,43	0,22	0,57
Autor	0	0	0	0	0,22	0,22
Autor + Cuerpo	0,09	0,57	0,52	0,52	0,22	0,57
Cuerpo	0,09	0,57	0,52	0,39	0,04	0,57

Tabla 5-2: Resultados para la temática de La Cumbre del Clima 2019 en el periódico El País

Fecha	Enlace de la noticia
2019-12-02 20:16	https://elpais.com/sociedad/2019/12/02/actualidad/1575268228_449028.html
2019-12-02 16:41	https://elpais.com/sociedad/2019/12/01/actualidad/1575216281_908130.html
2019-12-02 16:14	https://elpais.com/sociedad/2019/12/02/actualidad/1575284241_519772.html

2019-12-02 14:30	https://elpais.com/internacional/2019/12/02/mundo_global/1575287137_755469.html
2019-12-02 10:26	https://elpais.com/sociedad/2019/12/01/actualidad/1575226197_240379.html
2019-12-02 08:16	https://elpais.com/sociedad/2019/12/01/actualidad/1575197772_341810.html
2019-12-02 07:57	https://elpais.com/sociedad/2019/11/30/actualidad/1575139278_315997.html
2019-12-02 07:57	https://elpais.com/sociedad/2019/11/29/actualidad/1575022809_157378.html
2019-12-01 23:00	https://elpais.com/elpais/2019/12/01/opinion/1575217702_776463.html
2019-11-30 22:53	https://elpais.com/sociedad/2019/11/30/actualidad/1575127687_241694.html
2019-11-30 13:51	https://elpais.com/elpais/2019/11/30/album/1575114080_862435.html
2019-11-30 12:13	https://elpais.com/sociedad/2019/11/29/actualidad/1575060055_430396.html
2019-11-30 07:22	https://elpais.com/sociedad/2019/11/29/actualidad/1575052843_713181.html
2019-11-28 15:53	https://elpais.com/sociedad/2019/11/28/actualidad/1574934320_885860.html
2019-11-27 13:17	https://elpais.com/sociedad/2019/11/13/actualidad/1573642954_149954.html
2019-11-27 10:48	https://elpais.com/sociedad/2019/11/24/actualidad/1574621327_638200.html
2019-11-26 20:00	https://elpais.com/sociedad/2019/11/26/actualidad/1574791719_143580.html
2019-10-20 15:06	https://elpais.com/politica/2019/10/20/actualidad/1571576675_429915.html
2019-10-20 07:20	https://elpais.com/politica/2019/10/19/actualidad/1571492029_312414.html
2019-03-24 21:32	https://elpais.com/internacional/2019/03/22/actualidad/1553255888_370103.html
2019-03-22 13:54	https://elpais.com/internacional/2019/03/21/actualidad/1553156753_130649.html
2019-03-19 16:31	https://elpais.com/internacional/2019/03/19/actualidad/1552983067_837330.html
2019-02-18 09:16	https://elpais.com/ccaa/2019/02/17/catalunya/1550430376_996364.html
2019-02-02 18:03	https://elpais.com/ccaa/2019/02/02/catalunya/1549109317_249315.html

Tabla 5-3: Historial de noticias resultante para la temática de La Cumbre del Clima 2019 en el periódico El País

MEDIA DE LOS RESULTADOS	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas	
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	
Titular	0,38	0,48	0,48	0,46	0,2	0,5	0,42
Titular + Keywords	0,5	0,7	0,59	0,69	0,26	0,72	0,58
Titular + Resumen	0,32	0,5	0,46	0,49	0,23	0,51	0,42
Titular + Autor	0,36	0,49	0,47	0,49	0,21	0,49	0,42
Titular + Cuerpo	0,32	0,51	0,49	0,59	0,22	0,5	0,44
Keywords	0,52	0,7	0,61	0,7	0,25	0,73	0,59
Keywords + Resumen	0,46	0,69	0,51	0,63	0,23	0,7	0,54
Keywords + Autor	0,51	0,69	0,56	0,68	0,24	0,68	0,56
Keywords + Cuerpo	0,33	0,55	0,5	0,6	0,25	0,54	0,46
Resumen	0,21	0,3	0,28	0,3	0,23	0,33	0,28
Resumen + Autor	0,21	0,34	0,32	0,33	0,16	0,35	0,29
Resumen + Cuerpo	0,31	0,5	0,49	0,56	0,23	0,49	0,43
Autor	0,26	0,22	0,26	0,22	0,21	0,45	0,27
Autor + Cuerpo	0,3	0,51	0,48	0,56	0,23	0,5	0,43
Cuerpo	0,3	0,5	0,48	0,55	0,21	0,5	0,42
	0,35	0,51	0,47	0,52	0,22	0,53	

Tabla 5-4: Resultados medios obtenidos entre todos los experimentos

6 Conclusiones y trabajo futuro

6.1 Conclusiones

Se ha visto a lo largo del documento que es posible alcanzar el objetivo de crear un historial de noticias partiendo de una, con herramientas informáticas y algoritmos de comparación de textos. Esto ha supuesto mucho trabajo ya que una parte importante para alcanzar el objetivo del proyecto reside en tener un conjunto de datos con el qué operar, cosa que ha desembocado en realizar un desarrollo adicional para contar con la información de las diferentes noticias.

Una vez cubierta la parte del conjunto de datos, se ha tenido que reflexionar sobre los algoritmos que se iban a utilizar para la similitud de textos y derivaciones de estos que podrían aportar mejores resultados, además de qué tipo de tratamiento iban a tener los textos y formas de combinar la información de las noticias para así poder obtener los mejores resultados posibles.

Este TFG concluye aquí, donde se ha desarrollado una idea principal sólida que ha cumplido todos los objetivos propuestos. No obstante, quedan muchos experimentos por realizar, como un estudio donde se observen los resultados obtenidos realizando un tratamiento previo de los textos o no, probar el programa de creación de historiales con un dataset donde no todas las noticias estén tematizadas, etc. Por tanto, este proyecto puede servir de entrada o apoyo a otros que quieran investigar sobre esta rama de conocimiento.

Ahora, dando una conclusión un poco más personal sobre el proyecto, quiero remarcar que tanto mi tutor como yo nos hemos compenetrado, a mi parecer, lo mejor posible ya que no solo íbamos aportando ideas para abordar los objetivos del proyecto, si no que siempre le intentamos dar una vuelta de complejidad para que se viesen diferentes puntos de vista. La realización de este proyecto, aunque gratificante, ha sido complicada ya que una parte se ha realizado durante la vigencia del Estado de Alarma decretado por el Gobierno Español, lo cual nos hizo replantearnos rutinas y métodos de trabajo al no poder reunirnos presencialmente. Aun con todo esto, me siento muy orgulloso de haber realizado este proyecto ya que supone un broche final a mi carrera al haber podido trasladar gran parte de los conocimientos adquiridos durante estos últimos años a este trabajo.

6.2 Trabajo futuro

Este proyecto ha supuesto un gran trabajo y aunque se han querido abordar muchas ideas, aún nos queda mucho trabajo por realizar, sobre todo en tareas de optimización y actualización de la aplicación.

Algunas ideas para que mejore el proyecto y se estudien un poco más otras vías de comparación de textos son:

- **Realizar revisiones de las Webs de los periódicos.** Este es el punto más importante de todos. Como bien se sabe, una Web siempre está cambiando, además aquellas que tienen que ir reinventándose para que la apariencia atraiga a la mayor cantidad de usuarios posible. Por ello, cada cierto tiempo, se debería de hacer una prueba para ver si se está extrayendo la información de las Webs de forma correcta. Un ejemplo de esto es que, hace poco tiempo, la Web de **El País**, ha cambiado el dominio de las páginas a las que se navega desde su **hemeroteca** de ***"elpais.com"*** a ***"static.elpais.com"***, o la forma de la expresión **Xpath** a las noticias. Estos cambios podrían no englobar a toda la Web, si no solo a las páginas realizadas a partir de cierta fecha, por lo que supondría un trabajo adicional diferenciar entre épocas.
- **Creación de una base de datos para el guardado de noticias.** Ya que en este proyecto se ha optado por guardar la información de las noticias en ficheros JSON,

pensamos que crear una base de datos para optimizar el espacio de la información sería interesante debido a que además se podría modificar el módulo de creación de historiales, pudiendo mejorar su rendimiento a la hora de trabajar con las noticias.

- **Creación de una base de datos para el almacenamiento de resultados.** Ya que actualmente los resultados residen en ficheros de texto, pensamos que guardar esta información en una base de datos ayudaría a visualizar en mejor medida los resultados, así como poder realizar diagramas de resultados con mayor sencillez.
- **Estudiar otro tipo de representación de documentos y funciones de similitud.** En este proyecto solo se han estudiado 4, con 2 más que son variaciones de algunas de estas, pero aún quedan muchas más. Algunos ejemplos serían otros tipos de **Word Embeddings** que se citan en el apartado **2.3.3.1**, un estudio que pueda extraer las palabras clave de los textos o una mejora del estudio de noticias por fecha que se ha realizado en este proyecto entre muchas otras ideas.
- **Realizar un estudio de ablación de las características de las noticias.** Con este tipo de estudio se podría conseguir la mejor combinación de atributos para una función de similitud. Este estudio realizaría pruebas escogiendo un algoritmo el cual parta realizando la similitud de entre las noticias cogiendo todos sus atributos, de esta manera, por etapas, se van extrayendo aquellos atributos que peores resultados den, quedándose solo con la combinación de los atributos que más información aporten.

Referencias

[1] Dickinson, D. "*'Fake News' Challenges Audiences To Tell Fact From Fiction*", 2018. Disponible en: https://news.un.org/en/audio/2018/05/1008682. [Último acceso el 08-Jul-2019].

[2] Real Academia Española. Documentalista: Diccionario de la lengua española, 2020. Disponible en: https://dle.rae.es/documentalista. [Último acceso el 08-Jul-2019].

[3] El País. EL PAÍS: el periódico global, 2020. Disponible en: https://elpais.com/. [Último acceso el 08-Jul-2019].

[4] El Mundo. EL MUNDO - Diario online l, 2020. Disponible en: https://www.elmundo.es/. [Último acceso el 08-Jul-2019].

[5] 20 Minutos. El medio social - Última hora, local, España y el mundo, 2020. Disponible en: https://www.20minutos.es/. [Último acceso el 08-Jul-2019].

[6] El Confidencial. El Confidencial - El diario de los lectores influyentes, 2020. Disponible en: https://www.elconfidencial.com/. [Último acceso el 08-Jul-2019].

[7] Marca. MARCA - Diario online l, 2020. Disponible en: https://www.marca.com/. [Último acceso el 08-Jul-2019].

[8] Ryte. ¿Qué es un crawler o rastreador? 2020. Disponible en: https://es.ryte.com/wiki/Crawler. [Último acceso el 08-Jul-2019].

[9] Malik, S. K. y Rizvi, S. "Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation", 2011 International Conference on Computational Intelligence and Communication Networks, Gwalior, 2011.

[10] GRID - Digital Solutions. El web scraping: qué es, aplicaciones y consecuencias, 2019. Disponible en: https://www.grid.cl/blog/el-web-scraping-que-es-aplicaciones-y-consecuencias/. [Último acceso el 08-Jul-2019].

[11] Refsnes Data. CSS Selector Reference, 2020. Disponible en: https://www.w3schools.com/cssref/css_selectors.asp. [Último acceso el 08-Jul-2019].

[12] Refsnes Data. XPath Syntax, 2020. Disponible en: https://www.w3schools.com/xml/xpath_syntax.asp. [Último acceso el 08-Jul-2019].

[13] Scrapy Developers. Scrapy at a glance, 2020. Disponible en: https://docs.scrapy.org/en/latest/intro/overview.html. [Último acceso el 08-Jul-2019].

[14] Richardson, L. Beautiful Soup Documentation, 2020. Disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Último acceso el 08-Jul-2019].

[15] Selenium. The Selenium Browser Automation Project, 2020. Disponible en: https://www.selenium.dev/documentation/en/. [Último acceso el 08-Jul-2019].

[16] Pavalam, S. M., Kashmir Raja, S. V., Akorli, F. K. y Jawahar, M. "A Survey of Web Crawler Algorithms", 2011.

[17] Edwards, J., Mccurley, K. y Tomlin, J. "An Adaptive Model for Optimizing Performance of an Incremental Web Crawler", 2001.

[18] SAS Institute Inc. Qué es el Procesamiento de Lenguaje Natural - Natural Language Processing? 2018. Disponible en: https://www.sas.com/es_es/insights/analytics/what-is-natural-language-processing-nlp.html. [Último acceso el 08-Jul-2019].

[19] Perone, C. S. Machine Learning :: Cosine Similarity for Vector Space Models (Part III), Diciembre 09, 2013. Disponible en: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/. [Último acceso el 08-Jul-2019].

[20]Schnabel, T., Labutov, I., Mimno, D. y Joachims, T. “Evaluation methods for unsupervised word embeddings”. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.
[21]Mikolov, T., Chen, K., Corrado, G. y Dean, J. “Efficient Estimation of Word Representations in Vector Space”, 2013.
[22]Mikolov, T., Sutskever, I., Chen, K., Corrado, G. y Dean, J. “Distributed Representations of Words and Phrases and their Compositionality”, 2013.
[23]Facebook Inc. fastText, 2020. Disponible en: https://fasttext.cc/. [Último acceso el 08-Jul-2019].
[24]Pennington, J., Socher, R. y Manning, C. D. GloVe: Global Vectors for Word Representation, 2014. Disponible en: https://nlp.stanford.edu/projects/glove/.
[25]Pérez-Montoro Gutiérrez, M. "Sistemas de gestión de contenidos en la gestión del conocimiento", 2005.
[26]Ruiz, P., Dice, I., Ivan, Dice, L., LoMejordeWP, Dice, M., . . . Alonso, J. R. 45 Mejores Temas WordPress para Revistas y Periódicos 2020, 2020. Disponible en: https://www.lomejordewp.com/mejores-temas-wordpress-revistas-periodicos/. [Último acceso el 08-Jul-2019].
[27]WordPress. Crea un sitio web o blog gratuito, 2020. Disponible en: https://es.wordpress.com/. [Último acceso el 08-Jul-2019].
[28]Stark, N. S. “Motores De Búsqueda En Internet”, 2001.
[29]Real Academia Española. Hemeroteca: Diccionario de la lengua española, 2020. Disponible en: https://dle.rae.es/hemeroteca. [Último acceso el 08-Jul-2019].
[30]MongoDB. La base de datos líder del mercado para aplicaciones modernas, 2020. Disponible en: https://www.mongodb.com/es. [Último acceso el 08-Jul-2019].
[31]Google. Introducción a los archivos robots.txt - Ayuda de Search Console, 2020. Disponible en: https://support.google.com/webmasters/answer/6062608?hl=es. [Último acceso el 08-Jul-2019].
[32]Scrapy Developers. Architecture overview, 2020. Disponible en: https://docs.scrapy.org/en/latest/topics/architecture.html. [Último acceso el 08-Jul-2019].
[33]Manning, C. D., Raghavan, P. y Schütze, H. “*Introduction to information retrieval*”. Cambridge: Cambridge University Press. 2018.
[34]Explosion AI. SpaCy · Industrial-strength Natural Language Processing in Python, 2020. Disponible en: https://spacy.io/. [Último acceso el 08-Jul-2019].
[35]The SciPy community. NumPy v1.19 Manual, 2020. Disponible en: https://numpy.org/doc/stable/. [Último acceso el 08-Jul-2019].
[36]Scikit-Learn Developers. sklearn.metrics.pairwise.cosine_similarity — scikit-learn 0.23.1 documentation, 2020. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html. [Último acceso el 08-Jul-2019].
[37]Sadovsky, A. XPath Helper, 2015. Disponible en: https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl. [Último acceso el 08-Jul-2019].
[38]Scikit-Learn Developers. sklearn.metrics.classification_report — scikit-learn 0.23.1 documentation, 2020. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html. [Último acceso el 08-Jul-2019].
[39]DeepAI. F-Score, 2019. Disponible en: https://deepai.org/machine-learning-glossary-and-terms/f-score. [Último acceso el 08-Jul-2019].

[40]NLTK Project. Natural Language Toolkit, 2020. Disponible en: https://www.nltk.org/. [Último acceso el 08-Jul-2019].

Glosario

DOM	Document Object Model.
Dataset	Conjunto de datos.
HTML	HyperText Markup Language.
CSS	Cascading Style Sheet.
XML	Extensible Markup Language.
NLP	Natural Language Processing.
Crawler	Programa que navega por los diferentes enlaces de una Web de forma automática.
Scrapper	Programa que analiza el código de una página Web y extrae automáticamente información de esta.
Documentalista	Persona de la redacción de un periódico encargada de etiquetar las noticias en base a una o varias temáticas.
Hemeroteca digital	Web la cual aloja un archivo donde se guardan diarios y otras publicaciones periodísticas
URL	Uniform Resource Locator.
URI	Uniform Resource Identifier.

Anexos

A Funciones del módulo de construcción del historial de noticias

En este Anexo se expondrán las funciones que componen cada entidad del módulo de construcción del historial de noticias.

- **Extractor:**
  - **__init__**(): Constructor de la clase. Instancia el objeto ***nlp*** de **Spacy** para el tratamiento de objetos, además de inicializar variables para guardar las fechas en las que están comprendidas las noticias de la colección y la creación del diccionario con todos los datos de las noticias en memoria para que las búsquedas sean más rápidas.
  - **createDiccExtractor():** Crea el diccionario de noticias en memoria, la **clave** de este serán los **enlaces** y el **valor un diccionario** que contendrá como clave **los diferentes atributos** y como **valor sus respectivos textos**. Además, habrá otro diccionario que ordena las noticias por el tema que tratan estas, necesario para la evaluación de los resultados. Su **clave es el tema** y el **valor una lista de enlaces que tratan dicho tema**.
  - **getDataNoticiaMaster():** Función que devuelve el diccionario con los atributos de la noticia la cual queremos obtener su historial.
  - **getLinkNoticiaMaster():** Devuelve el enlace de la noticia que estamos estudiando.
  - **getDiccNoticiaEtiqueta():** Devuelve el diccionario con todos los temas que se tratan en la colección de noticias.
  - **getEtiquetasTema():** Devuelve el conjunto de temas tratados en la colección.
  - **getCountNoticiasConEtiqueta():** Devuelve el número de noticias que tratan un tema en específico.
  - **getLinksNoticiasAnalizar():** Devuelve una lista con todas las noticias de la colección menos la maestra o de estudio.
  - **getNumNoticiasAnalizar():** Devuelve el número de noticias en la colección sin contar la maestra.
  - **getDataNoticia():** Devuelve el diccionario de atributos de una noticia dado su link si y solo si la fecha de publicación de dicha noticia es anterior a la de la noticia maestra.
  - **getAtributosNoticia():** Dado el diccionario de una noticia y el atributo a extraer, se trata el texto del atributo en cuestión, realizando las técnicas propuestas en el punto **4.2.1.1** de la memoria, y se devuelve al usuario. Esta función tiene un modo en el que podemos saltarnos las tareas de tratamiento de textos si lo deseamos.
  - **create_diccRangoFechasNoticias():** Crea el diccionario que organiza las noticias por rangos de fecha y lo guarda en memoria. La forma en la que trabaja esta función es, primero crea los rangos de fechas de 4 horas en 4 horas y después coloca cada noticia en el rango que le corresponde.
  - **getVectorAtributo():** Devuelve el vector de características del texto de un atributo. Este vector de características se obtiene a partir del atributo ***vector*** del objeto **Documento** de un texto de **Spacy**.

- **getDoc():** Devuelve, en base a un texto, el objeto **Documento** de la librería ***Spacy***.
- **do_lemmatize():** Realiza la tarea de lematizar un texto. Para ello, se convierten todos los **tokens** o palabras de un texto a su *lema*, que este es aportado por el atributo *lemma_* del objeto **Token** de **Spacy**.
- **rm_stopWords():** Realiza la tarea de eliminar los *stopwords* de un texto. Para ello, se utiliza la librería **nltk**[40] *(Natural Language Toolkit)* que propone un conjunto de stopwords en español.
- **rm_punctuations():** Realiza la tarea de eliminación de signos de puntuación.
- **rm_blanks():** Realiza la tarea de eliminación de espacios en blanco que no sean necesarios, fruto de haber hecho el tratamiento de los textos.
- **openFile():** Abre el fichero donde se encuentran todos los datos de las noticias, devolviendo el puntero al fichero y su contenido.
- **closeFile():** Cierra el fichero donde se encuentran todos los datos de las noticias.

- **Similitud:**
  - **__init__():** Constructor de la clase Similitud. En él se inicializa el nombre de la similitud a crear. Dependiendo de la similitud se elegirá una función de similitud y se instanciaran diccionarios u otros objetos si fuera necesario. Dependiendo de cada caso podemos tener:
    - **Similitud coseno con vectores de características:** Únicamente nos hace falta instanciar la función de similitud que vamos a utilizar, en este caso **similitud_coseno_spacy()**.
    - **Similitud Jaccard:** Sucede lo mismo que en el caso anterior, aquí únicamente se asignará ***similitud_jaccard()*** como función de similitud a utilizar.
    - **Similitud coseno con vectorización tf-idf:** Aparte de inicializar la función de similitud, que en este caso será **similitud_coseno_links()**, se inicializan tres diccionarios. **El primero** con clave el identificador de una noticia, en este caso el enlace, y como valor otro diccionario que represente cada palabra de ese documento y la frecuencia de esta en él. **El segundo** representa cada palabra y el número de documentos en los que está presente. Por último, **el tercero** tiene como clave el identificador de la noticia y como valor el vector **numpy** que generaremos.
    - **Similitud coseno con vectorización BoW:** En este caso se inicializan las variables igual que en el caso anterior, excepto que aquí se elimina el **segundo diccionario** y se cambia por una **lista de palabras**, ya que en **BoW** solo se necesita conocer la frecuencia que tienen estas dentro de un texto.
  - **similitud_coseno_spacy():** Función que realiza la operación de similitud coseno entre dos textos a partir de la función **similarity(Doc)** del objeto documento de Spacy. Esta función internamente aplica la fórmula de similitud coseno a los dos vectores de características que da Spacy para un texto. Estos vectores se sacan por realizar la media entre todos los vectores de características de cada palabra contenida en el texto.
  - **similitud_jaccard():** Función que transforma cada documento a un Set de palabras y **realiza el cociente entre los elementos de la unión y los elementos de la intersección entre los dos conjuntos**.

- **similitud_coseno_links(): Función que calcula la similitud coseno entre dos documentos dados sus identificadores.** Aquí se tendrá que buscar en el objeto *self.dicc_doc_vector* los vectores de los documentos pedidos.
- **similitud_coseno_vecs(): Función que calcula la similitud coseno entre dos documentos dados sus vectores.**
- **add_doc_wFrec_entry(): Función que añade o actualiza una entrada al diccionario de la frecuencia de cada palabra por documento.** Este diccionario es el primero que comentamos en la función de **__init__()**, en el **caso de la similitud coseno**.
- **add_w_docsConW_entry(): Función que añade o actualiza una entrada al diccionario de frecuencias de las palabras en la colección.** Este diccionario es el segundo comentado en la función de **__init__()**, en el **caso de la similitud coseno con vectorización tf-idf**.
- **create_dicc_doc_tfidf(): Función que crea todos los vectores de documento en base a la vectorización tf-idf.**
- **add_doc_wFrec_entry_BoW(): Función idéntica a add_doc_wFrec_entry().** Está hecho de esta manera para optimizar los tiempos de ejecución y realizar las menos comparaciones posibles.
- **créate_vec_doc_BoW(): Función que crea todos los vectores de documento en base a la vectorización Bag of Words.**
- **getNombreSimilitud(): Devuelve el nombre de la similitud contenida en el objeto instanciado.**
- **getFuncionSimilitud(): Devuelve la función de similitud que utiliza el objeto instanciado.**
- **getSetPalabras(): Devuelve el Set de palabras de un documento dado.** Tener el Set es importante para el cálculo de Jaccard, ya que se necesitan las palabras que contiene el texto sin repeticiones.
- **getLinksNoticias(): Devuelve todos los enlaces de las noticias de la colección.**
- **getDocVector(): Devuelve el vector de una noticia dado su identificador.**

- **Procesador:**
  - **__init__()**: **Contructor de la clase Procesador.** En él se inicializan el **diccionario de resultados**, que tendrá de clave el identificador de la noticia y como clave la puntuación de similitud, y una **variable que guardará el score acumulado**.
  - **addResultado(): Función que añade un resultado al diccionario de resultados.**
  - **sortResultados(): Función que ordena el diccionario de resultados de mayor a menor puntuación.**
  - **getTopResultados(): Función que devuelve un String con todos los resultados obtenidos por cada noticia.** A esta función, si se desea, se le puede pasar el parámetro *top* que indicaría el número de noticias con mejor puntuación que nosotros queremos ver.
  - **__str__(): Función que proyecta las características del objeto Procesador,** como el número de noticias evaluadas y el score medio obtenido.
- **Main:**
  - **do_similitud_noCreacionVecs():** Función encargada de realizar el bucle de cálculo de similitudes entre noticias utilizando algoritmos de **similitud que no necesiten vectores** y con el **estudio simple**.

- **do_similitud_creacionVectores():** Función encargada de realizar el bucle de cálculo de similitudes entre noticias utilizando algoritmos de **similitud que necesiten realizar una vectorización de los documentos primero** y con el **estudio simple**.
- **do_similitud_noVecs_franjasHorarias():** Función encargada de realizar el bucle de cálculo de similitudes entre noticias utilizando algoritmos de similitud que **no necesiten vectores** y el **estudio de noticias por fecha**.
- **printResult(): Función que imprime los resultados obtenidos.**

B Manual de instalación

Para la instalación del proyecto primero hay que descargar todos los ficheros que los componen. Estos están subidos a **Github** en el siguiente repositorio:

https://github.com/andrescalvente997/historialNoticias

Una vez descargados los ficheros, tenemos:

- **/creacionDataset:** Directorio donde está guardada toda la funcionalidad referente al primer módulo de **recolección de noticias**.
- **/creacionHistorial:** Directorio donde está guardada toda la funcionalidad referente al segundo módulo de **construcción del historial de noticias**.
- **/resultados:** Directorio donde se guardan todos los ficheros de texto con los resultados obtenidos durante todo el proyecto.
- **/scripts:** Directorio que guarda diferentes scripts para la automatización de algunas tareas.
    - **Extraer_noticias.sh:** Script que extrae todas las noticias de todos los periódicos entre dos fechas preguntadas durante la ejecución.
    - **Extraer_noticias_tematica.py:** Script que en base a dar el directorio y el fichero donde están las noticias y un **tag**, extrae todas las noticias que contengan ese determinado **tag** en un fichero aparte, poniendo a estas noticias una temática propuesta por el usuario.
    - **Remove_noticias.sh:** Script que elimina todas los archivos de noticias contenidos en los directorios de los diferentes periódicos.

Una vista la estructura de los archivos que se han descargado y explicado algunos otros que no se han explicado antes en este documento, se tendrán que realizar los siguientes pasos para que funcionen todos los ejecutables:

- **Descargar la librería nltk y los stopwords que esta tiene:**

Para ello primero se debe descargar la librería con:

sudo pip3 install nltk

Una vez hecho, se tienen que descargar las *stopwords*:

python3
import nltk
nltk.download('stopwords')

- **Descargar la librería Spacy y su módulo de palabras en español:**

Para ello, se abrirá una terminal y se ejecutarán los siguientes dos comandos:

sudo pip3 install spacy
python3 -m Spacy download es_core_news_md

C Manual del programador

En este anexo se explicará cómo se debe hacer la ejecución de cada módulo. Los módulos que se pueden ejecutar son:

- **Módulo de recolección de noticias**

Para ejecutar el programa de extracción de noticias de una de las Webs seleccionadas, desde la raíz del proyecto debe de ejecutar por consola:

cd creacionDataset

Una vez allí, puede realizar varias ejecuciones. Por ejemplo, si se quiere extraer las noticas en un rango de fechas, se debe ejecutar el siguiente comando:

scrapy crawl <Spider> -a fechaIni="dd-MM-YYYY" -a fechaFin="dd-MM-YYYY"

La variable <***Spider***> hace referencia al nombre del spider, o lo que es lo mismo, de qué periódico se quieren extraer las noticias. Los nombres de las diferentes Spiders son:

- **Spider_ElPais**
- **Spider_ElMundo**
- **Spider_20Minutos**
- **Spider_ElConfidencial**
- **Spider_Marca**

También se puede realizar la extracción de noticias de un periódico de un mes entero. Esto se realizaría ejecutando:

scrapy crawl <Spider> -a anio="YYYY" -a mes="MM"

Además, también se puede realizar la ejecución de un día en concreto:

scrapy crawl <Spider> -a anio="YYYY" -a mes="MM" -a día="dd"

Por último, si se desea, se puede atribuir nombre al fichero de salida. Esto se realizará añadiendo el siguiente argumento a cualquiera de las ejecuciones anteriores:

-a strFile="Nombre_de_archivo.JSON"

- **Módulo de construcción del historial de noticias**

Para ejecutar el módulo de construcción de una noticia se debe primero entrar al directorio donde reside la funcionalidad. Esto se realizará ejecutando el siguiente comando:

cd creacionHistorial

Una vez aquí, para ejecutar el módulo se escribirá el siguiente comando en la consola:

python historialNoticias.py <Periódico> <URL_Noticia> <Temática> <Identificador>

Este módulo cogerá las noticias que estén situadas en el fichero **"/creacionDataset/crawlerPeriodicos/dataset_pruebas_ficheros/dataset_pruebas_PERIODICO.json".** El parámetro **URL_Noticia** hace referencia a la noticia con la que se harán

todas las similitudes y la más nueva en la cronología del historial. **Temática** hace referencia al tema principal de la noticia para hacer bien la evaluación de las noticias y construir bien los historiales, e **Identificador** únicamente servirá para poner nombre al fichero distinguiéndolo de los demás.

Una vez vistos los módulos principales, se explicará la ejecución de los distintos scripts creados:

- **Script de recolección de noticias**

Para ejecutar este script, desde la raíz se ejecuta el siguiente comando:

cd scripts

Una vez aquí, se escribirá el siguiente comando para ejecutar el script:

sh extraer_noticias.sh

Una vez lo ejecutemos, se nos preguntará por las fechas de inicio y fin de recolección. Se introducirán por input en el formato **"dd-MM-YYYY"** y el script ejecutará el módulo de recolección de noticias para todos los periódicos en ese rango de fechas especificado.

- **Script de borrado de noticias**

Este script borra todos los ficheros **.json** de los directorios donde se guardan las noticias de cada periódico. Los directorios que se ven involucrados son:

- **/datos_EL_PAIS**
- **/datos_EL_MUNDO**
- **/datos_20_MINUTOS**
- **/datos_EL_CONFIDENCIAL**
- **/datos_MARCA**

Para ello, desde la raíz se ejecutará el siguiente comando para entrar al directorio de scripts:

cd scripts

Una vez hecho, solo falta ejecutar el script de borrado:

sh remove_noticias.sh

Una vez hecho, el programa pregunta al usuario si realmente quiere borrar todos los ficheros. Está pregunta es únicamente de seguridad, si el usuario responde afirmativamente se borrarán los ficheros y en caso contrario terminará la ejecución.

- **Script para crear el dataset de pruebas**

Este script es el que se ha usado para extraer automáticamente todas las noticias que tratan sobre un tema en especial de entre todas las noticias extraídas en un rango de fecha. Este programa observa si en **tagsNoticia** existe una etiqueta determinada y si la tiene, extrae esa noticia y le asigna una temática. Para ejecutar este script se debe ir al directorio de scripts con el siguiente comando:

cd scripts

Una vez hecho, se arranca el programa de esta manera:

python extraer_noticias_tematica.py

Una vez hecho, el programa preguntará al usuario **el directorio y el archivo** donde están las noticias, después preguntará el **tag** que deben tener las noticias para que sean de esa temática y finalmente la **etiqueta de la temática** que el usuario quiera atribuir a esas noticias.

D Tablas de resultados por experimento

En este Anexo se muestran todas las tablas de todos los experimentos restantes realizados. Al final de este Anexo, se han calculado los resultados medios obtenidos por cada periódico.

JUICIO PROCÉS	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,34	0,25	0,36	0,25	0,22	0,22
Titular + Keywords	0,45	0,82	0,36	0,36	0,15	0,81
Titular + Resumen	0,23	0,31	0,36	0,28	0,15	0,35
Titular + Autor	0,4	0,26	0,39	0,26	0,25	0,19
Titular + Cuerpo	0,36	0,28	0,36	0,21	0,29	0,28
Keywords	0,48	0,83	0,43	0,57	0,15	0,82
Keywords + Resumen	0,35	0,78	0,44	0,6	0,18	0,76
Keywords + Autor	0,59	0,81	0,36	0,57	0,15	0,78
Keywords + Cuerpo	0,42	0,42	0,42	0,23	0,31	0,39
Resumen	0,19	0,35	0,5	0,35	0,15	0,32
Resumen + Autor	0,22	0,31	0,5	0,34	0,15	0,29
Resumen + Cuerpo	0,31	0,3	0,38	0,22	0,31	0,26
Autor	0,18	0,4	0,52	0,4	0,15	0,39
Autor + Cuerpo	0,38	0,27	0,35	0,22	0,29	0,25
Cuerpo	0,33	0,26	0,35	0,21	0,29	0,26

Tabla D-1: Resultados para la temática de El Juicio del Procés en el periódico El Confidencial

Fecha	Enlace
2019-10-20 12:27:00	https://www.elconfidencial.com/espana/cataluna/2019-10-20/quim-torra-vuelve-a-llamar-a-pedro-sanchez-sin-obtener-respuesta_2291844/
2019-10-19 19:59:00	https://www.elconfiden+E4:E51cial.com/espana/cataluna/2019-10-19/destrozos-transporte-publico-barcelona-disturbios_2291499/
2019-10-19 03:00:00	https://www.elconfidencial.com/empresas/2019-10-19/indiferencia-inversora-crisis-cataluna-inmobiliario_2289635/
2019-10-19 03:00:00	https://www.elconfidencial.com/elecciones-generales/2019-10-19/psoe-enquistamiento-tension-cataluna-castigo-pedro-sanchez-10n_2290628/
2019-10-19 03:00:00	https://www.elconfidencial.com/tecnologia/2019-10-19/tsunami-democratic-aepd-app-privacidad-proces-barcelona-sentencia_2289399/

2019-10-18 23:25:00	https://www.elconfidencial.com/espana/cataluna/2019-10-19/sentencia-proces-directo-hora-barcelona-huelga-cataluna-434_2287187/
2019-10-18 11:51:00	https://www.elconfidencial.com/espana/cataluna/2019-10-18/huelga-barcelona-proces-tiendas-manifestaciones-221_2289779/
2019-10-18 06:56:00	https://www.elconfidencial.com/espana/cataluna/2019-10-18/puigdemont-acude-fiscalia-nueva-euroorden-llarena-557_2289043/
2019-10-16 08:40:00	https://www.elconfidencial.com/espana/2019-10-16/gobierno-helicoptero-franco-exhumacion-supremp_2285459/
2019-10-15 15:42:00	https://www.elconfidencial.com/espana/cataluna/2019-10-15/sentencia-proces-torra-contradiicones-llamar-protesta-enviar-mossos_2284292/
2019-10-15 13:50:00	https://www.elconfidencial.com/tecnologia/2019-10-15/sentencia-fakes-imagenes-falsas-manifestaciones-656_2283279/
2019-10-15 12:02:00	https://www.elconfidencial.com/tecnologia/2019-10-15/app-tsunami-democratic-supremo-sentencia-proces-cataluna-015_2283628/
2019-10-15 10:45:00	https://www.elconfidencial.com/espana/cataluna/2019-10-15/sentencia-proces-erc-gabriel-rufian-pelea-985_2283591/
2019-10-15 03:00:00	https://www.elconfidencial.com/espana/2019-10-15/jueces-supremo-repudian-acusaciones-venganza-politicos-secesionistas_2282279/
2019-10-15 03:00:00	https://www.elconfidencial.com/espana/2019-10-15/sentencia-proces-euroorden-puigdemont_2282663/
2019-10-15 03:00:00	https://www.elconfidencial.com/espana/cataluna/2019-10-15/sentencia-proces-caos-aeropuerto-guardiola-eco-europa-142_2282963/
2019-10-14 20:53:00	https://www.elconfidencial.com/espana/2019-10-14/manifestaciones-sentencia-proces-video-cargas_2282984/
2019-10-14 18:21:00	https://www.elconfidencial.com/espana/cataluna/2019-10-14/sentencia-proces-valtonic-vox-girona-bandera-espana-763_2282864/
2019-10-14 17:23:00	https://www.elconfidencial.com/espana/cataluna/2019-10-14/sentencia-proces-tsunami-democratic-aeropuerto-prat_2282659/
2019-10-14 15:53:00	https://www.elconfidencial.com/espana/cataluna/2019-10-14/vuelos-cancelados-protestas-barcelona-sentencia_2282516/
2019-10-14 14:44:00	https://www.elconfidencial.com/espana/cataluna/2019-10-14/carles-puigdemont-reaccion-sentencia-proces_2282383/
2019-10-14 11:01:00	https://www.elconfidencial.com/espana/cataluna/2019-10-14/sentencia-proces-oriol-junqueras-justicia-venganza_2281872/
2019-10-13 03:00:00	https://www.elconfidencial.com/espana/2019-10-13/el-gobierno-preve-escenarios-fallo-apoyo-total-pp_2280512/
2019-10-13 03:00:00	https://www.elconfidencial.com/espana/2019-10-13/sentencia-proces-junqueras-condena-sedicion-522_2280215/
2019-10-12 07:41:00	https://www.elconfidencial.com/espana/2019-10-12/tribunal-supremo-condenara-sedicion-juicio-proces-746_2278464/
2019-10-12 03:00:00	https://www.elconfidencial.com/espana/2019-10-12/malestar-filtraciones-sentencia-amenaza-dilatar-tiempos-521_2278840/

2019-10-11 12:54:00	https://www.elconfidencial.com/espana/cataluna/2019-10-11/antidisturbios-barcelona-sentencia-proces-protestas_2278027/
2019-10-09 15:45:00	https://www.elconfidencial.com/espana/2019-10-09/siete-magistrados-audiencia-nacional-postulan-supremo_2274279/
2019-10-08 15:02:00	https://www.elconfidencial.com/espana/cataluna/2019-10-08/mossos-cataluna-protestas-seguridad-efectivos-391_2274148/
2019-10-08 14:47:00	https://www.elconfidencial.com/espana/2019-10-08/fiscalia-pide-fijar-la-vista-de-prorroga-de-prision-de-los-presos-del-proces_2274136/
2019-10-08 03:00:00	https://www.elconfidencial.com/espana/2019-10-08/tribunal-proces-penas-acusados-sentencia_2272348/
2019-10-07 03:00:00	https://www.elconfidencial.com/espana/cataluna/2019-10-07/manual-afrontar-sentencia-proces-calle-catalonia_2269871/
2019-10-05 17:47:00	https://www.elconfidencial.com/elecciones-generales/2019-10-05/sanchez-replica-a-rivera-que-pide-que-no-bloquee_2270516/
2019-10-05 10:47:00	https://www.elconfidencial.com/espana/cataluna/2019-10-05/scc-cataluna-no-necesita-marchas-en-sentencia-proces_2270264/
2019-10-04 17:45:00	https://www.elconfidencial.com/espana/2019-10-04/franco-agarran-prior-valle-dilatar-exhumacion-ts_2269596/
2019-10-04 03:00:00	https://www.elconfidencial.com/espana/cataluna/2019-10-04/puigdemont-impone-ley-pdecat-listas-10n_2268291/
2019-10-03 03:00:00	https://www.elconfidencial.com/elecciones-generales/2019-10-03/entrevista-pedro-sanchez-cataluna-independentismo-unidad_2266427/
2019-10-02 18:15:00	https://www.elconfidencial.com/espana/cataluna/2019-10-02/asi-prepara-respuesta-sentencia-proces-978_2266615/
2019-10-02 11:30:00	https://www.elconfidencial.com/elecciones-generales/2019-10-02/entrevista-pedro-sanchez-avance-ley-seguridad-nacional-155_2265500/
2019-04-30 03:00:00	https://www.elconfidencial.com/espana/cataluna/2019-04-30/puigdemont-tension-erc-independencia_1973266/
2019-04-28 19:17:00	https://www.elconfidencial.com/elecciones-generales/2019-04-28/elecciones-generales-resultados-cataluna_1970258/
2019-04-26 19:19:00	https://www.elconfidencial.com/elecciones-generales/2019-04-26/elecciones-rivera-ciudadanos-campanada_1967442/
2019-04-25 20:08:00	https://www.elconfidencial.com/elecciones-generales/2019-04-25/sede-psoe-madrid-atacada-afilio-pedro-sanchez_1964458/
2019-04-23 22:13:00	https://www.elconfidencial.com/elecciones-generales/2019-04-24/debate-electoral-pastor-valles-escaramuza_1958946/
2019-04-22 23:00:00	https://www.elconfidencial.com/elecciones-generales/2019-04-23/debate-electoral-2019-sanchez-casado-iglesias-rivera_1956210/
2019-04-21 15:56:00	https://www.elconfidencial.com/espana/cataluna/2019-04-21/elecciones-pp-alvarez-de-toledo-cataluna-sanchez-psoe_1953434/
2019-04-17 21:19:00	https://www.elconfidencial.com/elecciones-generales/2019-04-17/jec-jordi-sanchez-actos-electorales-carcel-elecciones-junqueras-romeva_1950886/

2019-04-15 14:26:00	https://www.elconfidencial.com/elecciones-generales/2019-04-15/grupo-whatsapp-pedro-sanchez-lona-calle-goya-ciudadanos_1945138/
2019-03-25 18:49:00	https://www.elconfidencial.com/espana/cataluna/2019-03-25/nueva-querella-torra-lazo-generalitat_1902950/
2019-03-25 15:53:00	https://www.elconfidencial.com/espana/2019-03-25/juicio-proces-guardia-civil-antidisturbios-20s_1902450/
2019-03-25 12:47:00	https://www.elconfidencial.com/espana/2019-03-25/juicio-proces-marchena-guardia-civil-evasivas_1901994/
2019-03-25 10:29:00	https://www.elconfidencial.com/espana/2019-03-25/juicio-proces-20s-guardia-civil-mossos-indignacion_1901182/
2019-03-24 04:00:00	https://www.elconfidencial.com/espana/2019-03-24/juicio-proces-preguntas-frikis-tribunal-supremo_1898374/
2019-03-22 04:00:00	https://www.elconfidencial.com/espana/2019-03-22/violencia-lagrimas-guardia-civil-desprecio-1-o_1896590/
2019-03-21 16:41:00	https://www.elconfidencial.com/espana/2019-03-21/juicio-proces-guardia-civil-referendum-emociona_1896670/
2019-03-21 04:00:00	https://www.elconfidencial.com/espana/2019-03-21/mossos-guardia-civil-vigilancia-companeros_1894266/
2019-03-20 19:54:00	https://www.elconfidencial.com/espana/cataluna/2019-03-20/torra-no-dara-la-orden-de-retirar-los-lazos-pero_1894510/
2019-03-20 09:29:00	https://www.elconfidencial.com/espana/2019-03-20/juicio-proces-quim-torra-guardia-civil-papeletas_1892642/
2019-03-20 04:00:00	https://www.elconfidencial.com/espana/2019-03-20/rabia-extrema-y-terror-total-los-guardia-civiles-hablan-de-violencia-descontrolada_1892346/
2019-03-20 04:00:00	https://www.elconfidencial.com/espana/2019-03-20/edmundo-bal-ciudadadanos-abogacia-estado-proces_1891146/
2019-03-19 19:56:00	https://www.elconfidencial.com/espana/2019-03-19/jose-maria-aznar-indultos-gobiernos-nacionalistas_1892254/
2019-03-19 13:12:00	https://www.elconfidencial.com/espana/2019-03-19/juicio-proces-marchena-alonso-martinez-defensas_1891346/
2019-03-19 10:51:00	https://www.elconfidencial.com/espana/2019-03-19/juicio-proces-guardia-civil-salvado-papeles-ventana_1890510/
2019-03-19 04:00:00	https://www.elconfidencial.com/espana/2019-03-19/tribunal-reorganiza-juicio-sesiones-campanas-electorales_1889614/
2019-03-18 04:00:00	https://www.elconfidencial.com/espana/2019-03-18/juicio-cataluna-supremo-guardias-civiles-registros_1887470/
2019-03-16 19:08:00	https://www.elconfidencial.com/espana/2019-03-16/manifestacion-independentista-madrid-asistencia_1886726/
2019-03-16 04:00:00	https://www.elconfidencial.com/espana/2019-03-16/manifestacion-independentista-madrid-recorrido_1885738/
2019-03-16 04:00:00	https://www.elconfidencial.com/espana/2019-03-16/de-heroe-a-villano-trapero-pasa-a-la-lista-negra-del-independentismo_1884858/

2019-03-15 09:58:00	https://www.elconfidencial.com/espana/2019-03-15/defensas-proces-queja-marchena-suplanto-acusaciones_1884002/
2019-03-14 10:52:00	https://www.elconfidencial.com/espana/2019-03-14/trapero-juicio-proces-marchena-supremo-mossos_1880854/
2019-03-13 11:11:00	https://www.elconfidencial.com/espana/2019-03-13/accion-exterior-diplocat-albert-royo-juicio-proces_1878554/
2019-02-20 20:36:00	https://www.elconfidencial.com/espana/2019-02-20/mundo-niega-que-desobedeciera-al-tc-en-un-tenso-interrogatorio-con-el-fiscal_1838418/
2019-02-20 18:48:00	https://www.elconfidencial.com/espana/2019-02-20/juicio-proces-borras-supremo-consellera-dui_1838170/
2019-02-20 17:55:00	https://www.elconfidencial.com/espana/2019-02-20/juicio-proces-bassa-niega-estrategia-conjunta-independencia-pactada_1837834/
2019-02-20 11:22:00	https://www.elconfidencial.com/espana/2019-02-20/juicio-proces-josep-rull-declaracion-supremo_1836578/
2019-02-19 19:49:00	https://www.elconfidencial.com/espana/2019-02-19/leccion-derecho-turull-romeva-constitucion-autodeterminacion_1835710/
2019-02-19 17:25:00	https://www.elconfidencial.com/espana/2019-02-19/juicio-proces-raul-romeva-interrogatorio-vox_1835318/
2019-02-19 11:42:00	https://www.elconfidencial.com/espana/2019-02-19/rajoy-y-santamaria-declararan-como-testigos-la-proxima-semana_1834678/
2019-02-19 10:18:00	https://www.elconfidencial.com/espana/2019-02-19/juicio-proces-turull-declaracion-fiscalia_1834162/
2019-02-19 04:00:00	https://www.elconfidencial.com/espana/2019-02-19/contraobservadores-juicio-proces-juristas-defiende-constitucion_1832282/
2019-02-19 04:00:00	https://www.elconfidencial.com/espana/2019-02-19/juicio-proces-turull-via-forn-contestara-fiscalia_1832706/
2019-02-18 04:00:00	https://www.elconfidencial.com/espana/2019-02-18/supremo-rajoy-politicos-campana-electoral-vox_1829362/
2019-02-17 04:00:00	https://www.elconfidencial.com/cultura/2019-02-17/lluis-companys-juicio-gobierno-catalan_1825770/
2019-02-16 04:00:00	https://www.elconfidencial.com/espana/2019-02-16/espana-turca-espana-sueca-juicio-proces_1827654/
2019-02-16 04:00:00	https://www.elconfidencial.com/espana/2019-02-16/juicio-proces-sanchez-embajadas-independentismo_1828798/
2019-02-15 21:02:00	https://www.elconfidencial.com/espana/2019-02-15/la-ministra-de-justicia-se-fotografia-en-barcelona-con-uno-de-los-abogados-del-proces_1829350/
2019-02-14 17:21:00	https://www.elconfidencial.com/espana/2019-02-14/directo-juicio-proces-junqueras-supremo_1823986/
2019-02-14 04:00:00	https://www.elconfidencial.com/espana/2019-02-14/juicio-proces-tribunal-admitira-algunas-peticiones-defensas-ampliara-testifical_1824378/
2019-02-13 23:00:00	https://www.elconfidencial.com/espana/2019-02-14/juicio-proces-torra-dibujo-consellera-goirizelaia_1824394/

2019-02-13 09:04:00	https://www.elconfidencial.com/espana/cataluna/2019-02-13/torra-sanchez-21puntos-juicio-proces_1822378/
2019-02-13 04:00:00	https://www.elconfidencial.com/espana/2019-02-13/mundo-junqueras-torra-juicio-proces-bano_1822062/
2019-02-12 16:53:00	https://www.elconfidencial.com/espana/cataluna/2019-02-12/cdr-juicio-proces-cataluna-protestas_1821422/
2019-02-12 09:56:00	https://www.elconfidencial.com/espana/2019-02-12/juicio-proces-directo-supremo_1818422/
2019-02-11 04:00:00	https://www.elconfidencial.com/espana/2019-02-11/vox-juicio-proces-rajoy-independentistas_1814574/
2019-02-11 04:00:00	https://www.elconfidencial.com/espana/2019-02-11/el-tribunal-del-proces-los-siete-magistrados-que-juzgaran-a-los-politicos-catalanes_1812202/
2019-02-10 04:03:00	https://www.elconfidencial.com/espana/2019-02-10/la-linea-de-defensa-del-proces-siete-pretigiosos-abogados-catalanes_1814614/
2019-02-02 12:40:00	https://www.elconfidencial.com/espana/2019-02-02/claves-juicio-proces-acusados-testigos-guardia-civil-auto_1799990/

Tabla D-2: Historial de noticias resultante para la temática de El Juicio del Procés en el periódico El Confidencial

BREXIT	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,08	0,19	0,27	0,19	0,1	0,19
Titular + Keywords	0,5	0,89	0,98	0,87	0,1	0,89
Titular + Resumen	0,23	0,24	0,26	0,16	0,18	0,24
Titular + Autor	0,1	0,44	0,45	0,44	0,05	0,44
Titular + Cuerpo	0,23	0,4	0,34	0,29	0,24	0,4
Keywords	0,63	0,87	0,95	0,97	0,06	0,97
Keywords + Resumen	0,66	0,89	0,26	0,87	0,11	0,89
Keywords + Autor	0,29	0,87	0,84	0,94	0	0,94
Keywords + Cuerpo	0,24	0,42	0,35	0,34	0,26	0,42
Resumen	0,27	0,11	0,08	0,11	0,24	0,1
Resumen + Autor	0,18	0,42	0,44	0,42	0,05	0,42
Resumen + Cuerpo	0,24	0,39	0,35	0,27	0,23	0,39
Autor	0,52	0,47	0,44	0,47	0,02	0,48
Autor + Cuerpo	0,21	0,4	0,35	0,29	0,21	0,4
Cuerpo	0,23	0,4	0,35	0,29	0,21	0,4

Tabla D-3: Resultados para la temática de El Brexit en el periódico El Mundo

Fecha	Enlace
2020-02-09 18:46:09	https://www.elmundo.es/internacional/2020/02/09/5e405358fdddff0f278b45bd.html
2020-02-03 11:14:34	https://www.elmundo.es/internacional/2020/02/03/5e380088fc6c83f4518b45b6.html
2020-01-31 18:51:32	https://www.elmundo.es/internacional/2020/01/31/5e347403fdddff8b5b8b4639.html
2020-01-31 17:46:39	https://www.elmundo.es/internacional/2020/01/31/5e346611fc6c83c0198b4577.html
2020-01-31 12:32:12	https://www.elmundo.es/internacional/2020/01/31/5e341bf9fdddff831f8b4674.html
2020-01-31 11:39:07	https://www.elmundo.es/internacional/2020/01/31/5e3411cafdddff32148b4611.html
2020-01-31 07:57:10	https://www.elmundo.es/cultura/literatura/2020/01/31/5e3054b8fc6c8330438b456f.html

2020-01-31 01:05:07	https://www.elmundo.es/opinion/editorial/2020/01/31/5e331537fdddff2a348b457b.html
2020-01-31 01:02:40	https://www.elmundo.es/internacional/2020/01/31/5e333a5efdddff4c268b45b7.html
2020-01-31 00:59:10	https://www.elmundo.es/internacional/2020/01/31/5e32fea5fdddfff2378b458d.html
2020-01-30 20:43:33	https://www.elmundo.es/internacional/2020/01/30/5e330d2221efa08e128b4639.html
2020-01-25 12:36:27	https://www.elmundo.es/internacional/2020/01/25/5e2c355cfdddff1e488b463e.html
2020-01-22 21:08:34	https://www.elmundo.es/internacional/2020/01/22/5e28b62a21efa0c8768b45e3.html
2019-10-20 17:03:35	https://www.elmundo.es/internacional/2019/10/20/5dac9356fc6c834b678b46a1.html
2019-10-20 00:47:29	https://www.elmundo.es/internacional/2019/10/20/5daaf2b3fc6c83a0658b45db.html
2019-10-19 14:42:24	https://www.elmundo.es/internacional/2019/10/19/5dab2057fc6c83fb5e8b45ea.html
2019-10-19 13:29:12	https://www.elmundo.es/internacional/2019/10/19/5dab0e19fc6c8361718b45c3.html
2019-10-19 09:26:22	https://www.elmundo.es/internacional/2019/10/19/5daad5f2fc6c83d8658b45a8.html
2019-10-18 00:07:40	https://www.elmundo.es/opinion/2019/10/18/5da89c31fdddff55988b45e5.html
2019-10-18 00:01:34	https://www.elmundo.es/internacional/2019/10/18/5da8b50b21efa008368b45d9.html
2019-10-17 18:46:56	https://www.elmundo.es/internacional/2019/10/17/5da8b54b21efa07a6e8b45da.html
2019-10-17 14:38:33	https://www.elmundo.es/internacional/2019/10/17/5da87c97fdddffd5528b45c1.html
2019-10-17 09:43:27	https://www.elmundo.es/internacional/2019/10/17/5da615ff21efa0c3748b462b.html
2019-10-17 09:42:03	https://www.elmundo.es/internacional/2019/10/17/5da5f9e9fdddff7e8f8b457a.html
2019-10-17 09:36:08	https://www.elmundo.es/internacional/2019/10/17/5da835b921efa0a4288b4697.html
2019-10-17 08:07:23	https://www.elmundo.es/internacional/2019/10/17/5da82128fdddff114b8b45a3.html
2019-10-17 00:09:39	https://www.elmundo.es/espana/2019/10/17/5da74f3021efa0860d8b466d.html
2019-10-16 18:04:18	https://www.elmundo.es/internacional/2019/10/16/5da75905fdddff12808b4570.html

2019-10-16 14:20:55	https://www.elmundo.es/internacional/2019/10/16/5da7273221efa0c3748b469f.html
2019-10-15 18:28:06	https://www.elmundo.es/internacional/2019/10/15/5da60fa521efa0bc6e8b45f6.html
2019-10-15 11:58:50	https://www.elmundo.es/internacional/2019/10/15/5da48062fdddff88898b46b6.html
2019-10-15 07:15:15	https://www.elmundo.es/internacional/2019/10/15/5da571dafc6c83677c8b4601.html
2019-10-12 18:47:22	https://www.elmundo.es/espana/2019/10/12/5da21fabfc6c83f2588b4687.html
2019-10-11 12:55:00	https://www.elmundo.es/internacional/2019/10/11/5da0798e21efa0ad238b45b1.html
2019-10-10 17:04:34	https://www.elmundo.es/internacional/2019/10/10/5d9f6492fdddffa4378b4768.html
2019-10-09 00:03:49	https://www.elmundo.es/cronica/2019/10/09/5d91e44bfc6c83d7238b46a1.html
2019-10-08 11:43:46	https://www.elmundo.es/internacional/2019/10/08/5d9c766021efa0723c8b466e.html
2019-10-07 12:41:10	https://www.elmundo.es/internacional/2019/10/07/5d9b325421efa0cf268b4633.html
2019-10-07 00:19:39	https://www.elmundo.es/internacional/2019/10/07/5d9a0d02fdddffba7c8b4594.html
2019-10-06 09:04:46	https://www.elmundo.es/internacional/2019/10/06/5d99ae1efdddff069f8b4575.html
2019-10-03 15:10:32	https://www.elmundo.es/internacional/2019/10/03/5d960d2021efa0db368b45bf.html
2019-10-03 11:39:55	https://www.elmundo.es/internacional/2019/10/03/5d95d620fc6c8371488b459f.html
2019-10-02 20:28:51	https://www.elmundo.es/internacional/2019/10/02/5d950876fc6c83eb3b8b456f.html
2019-10-02 15:37:24	https://www.elmundo.es/internacional/2019/10/02/5d94c3a5fc6c8355538b465f.html
2019-10-02 11:32:55	https://www.elmundo.es/internacional/2019/10/02/5d948ad0fc6c830a6f8b4629.html
2019-10-02 00:18:30	https://www.elmundo.es/internacional/2019/10/02/5d93ecc6fc6c83a25e8b462c.html
2019-10-01 12:34:15	https://www.elmundo.es/internacional/2019/10/01/5d934720fc6c83d2628b45b9.html
2019-03-25 16:45:00	https://www.elmundo.es/internacional/2019/03/25/5c98f985fc6c8395208b4572.html
2019-03-24 18:01:35	https://www.elmundo.es/internacional/2019/03/24/5c97c5e0fc6c83ab0b8b466d.html

2019-03-24 09:13:08	https://www.elmundo.es/internacional/2019/03/24/5c974953fdddff3c5a8b4665.html
2019-03-24 01:13:39	https://www.elmundo.es/internacional/2019/03/24/5c969977fc6c832c2f8b4637.html
2019-03-23 01:22:22	https://www.elmundo.es/opinion/2019/03/23/5c95200021efa0ed5e8b468f.html
2019-03-22 18:37:57	https://www.elmundo.es/internacional/2019/03/22/5c952b76fc6c836f158b45b1.html
2019-03-21 21:18:52	https://www.elmundo.es/internacional/2019/03/21/5c93ff3821efa0f14b8b459a.html
2019-03-21 15:10:13	https://www.elmundo.es/internacional/2019/03/21/5c93a8f7fc6c83741d8b45fa.html
2019-03-21 13:02:01	https://www.elmundo.es/economia/ahorro-y-consumo/2019/03/21/5c938908fdddff08708b471f.html
2019-03-20 20:58:05	https://www.elmundo.es/internacional/2019/03/20/5c92a0a4fc6c8366048b4576.html
2019-03-20 16:30:27	https://www.elmundo.es/internacional/2019/03/20/5c92677321efa04d698b4692.html
2019-03-20 10:11:47	https://www.elmundo.es/internacional/2019/03/20/5c9211be21efa07a618b4584.html
2019-03-15 01:05:58	https://www.elmundo.es/internacional/2019/03/15/5c8aac90fdddff6b178b45a4.html
2019-03-14 18:25:02	https://www.elmundo.es/internacional/2019/03/14/5c8a91cefdddffbe2c8b45f4.html
2019-03-14 01:07:47	https://www.elmundo.es/opinion/2019/03/14/5c8913e8fc6c8307218b4573.html
2019-03-13 19:22:09	https://www.elmundo.es/internacional/2019/03/13/5c89555421efa054018b46d4.html

Tabla D-4: Historial de noticias resultante para la temática de El Brexit en el periódico El Mundo

NÔTRE DAME	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,75	0,62	0,5	0,62	0,4	0,65
Titular + Keywords	0,35	0,65	0,57	0,78	0,3	0,65
Titular + Resumen	0,28	0,62	0,48	0,6	0,2	0,62
Titular + Autor	0,7	0,62	0,48	0,62	0,45	0,65
Titular + Cuerpo	0,48	0,57	0,68	0,82	0,33	0,57
Keywords	0,2	0,85	0,53	0,65	0,2	0,85
Keywords + Resumen	0,17	0,8	0,53	0,55	0,17	0,8
Keywords + Autor	0,23	0,85	0,55	0,65	0,25	0,85
Keywords + Cuerpo	0,42	0,6	0,62	0,8	0,38	0,6
Resumen	0,15	0,48	0,45	0,45	0,07	0,5
Resumen + Autor	0,12	0,53	0,5	0,5	0,1	0,53
Resumen + Cuerpo	0,4	0,57	0,65	0,78	0,3	0,57
Autor	0,48	0,38	0,33	0,38	0,35	0,8
Autor + Cuerpo	0,42	0,55	0,7	0,78	0,33	0,55
Cuerpo	0,42	0,55	0,68	0,78	0,33	0,55

Tabla D-5: Resultados para la temática de El Incendio de la Catedral de Nôtre Dame en el periódico El País

Fecha	Enlace
2019-04-29 20:43:47	https://elpais.com/cultura/2019/04/27/actualidad/1556380153_549141.html
2019-04-28 15:06:14	https://elpais.com/elpais/2019/04/26/icon_design/1556296401_058318.html
2019-04-25 21:15:45	https://elpais.com/elpais/2019/04/24/paco_nadal/1556134245_307614.html
2019-04-24 19:14:53	https://elpais.com/cultura/2019/04/24/actualidad/1556107937_447747.html
2019-04-21 22:00:00	https://elpais.com/elpais/2019/04/21/opinion/1555863321_213331.html
2019-04-21 13:15:04	https://elpais.com/cultura/2019/04/20/actualidad/1555774594_554018.html
2019-04-20 22:00:00	https://elpais.com/elpais/2019/04/20/opinion/1555768422_928660.html

2019-04-20 22:00:00	https://elpais.com/elpais/2019/04/20/opinion/1555780226_545616.html
2019-04-20 19:11:15	https://elpais.com/internacional/2019/04/20/actualidad/1555753358_727469.html
2019-04-19 18:21:55	https://elpais.com/elpais/2019/04/19/album/1555676976_292348.html
2019-04-19 16:04:23	https://elpais.com/cultura/2019/04/18/actualidad/1555606604_970892.html
2019-04-18 21:28:03	https://elpais.com/cultura/2019/04/17/actualidad/1555531561_800883.html
2019-04-18 07:08:45	https://elpais.com/cultura/2019/04/17/actualidad/1555499782_313410.html
2019-04-18 07:07:23	https://elpais.com/cultura/2019/04/16/actualidad/1555432161_255893.html
2019-04-18 06:01:37	https://elpais.com/cultura/2019/04/17/actualidad/1555526261_870552.html
2019-04-17 22:00:57	https://elpais.com/elpais/2019/04/17/gente/1555518252_651785.html
2019-04-17 14:53:43	https://elpais.com/cultura/2019/04/16/actualidad/1555408785_837467.html
2019-04-17 10:56:44	https://elpais.com/elpais/2019/04/16/album/1555401446_249882.html
2019-04-17 06:38:46	https://elpais.com/internacional/2019/04/16/actualidad/1555401757_585012.html
2019-04-17 06:12:18	https://elpais.com/cultura/2019/04/16/actualidad/1555431502_955651.html
2019-04-16 22:00:27	https://elpais.com/elpais/2019/04/16/opinion/1555425243_571004.html
2019-04-16 22:00:00	https://elpais.com/elpais/2019/04/16/opinion/1555430561_275098.html
2019-04-16 21:35:42	https://elpais.com/cultura/2019/04/15/actualidad/1555356902_708073.html
2019-04-16 21:30:40	https://elpais.com/sociedad/2019/04/16/actualidad/1555435892_325415.html
2019-04-16 19:22:12	https://elpais.com/cultura/2019/04/16/actualidad/1555412193_499224.html
2019-04-16 18:52:30	https://elpais.com/cultura/2019/04/16/actualidad/1555397858_382444.html
2019-04-16 18:30:36	https://elpais.com/elpais/2019/04/16/videos/1555431689_133615.html
2019-04-16 15:55:54	https://elpais.com/elpais/2019/04/16/videos/1555427557_044193.html

2019-04-16 12:01:08	https://elpais.com/internacional/2019/04/15/estados_unidos/1555356066_396871.html
2019-04-16 12:00:45	https://elpais.com/internacional/2019/04/15/actualidad/1555328528_882043.html
2019-04-16 11:15:47	https://elpais.com/cultura/2019/04/16/actualidad/1555408947_874026.html
2019-04-16 11:02:56	https://elpais.com/cultura/2019/04/15/actualidad/1555359142_362044.html
2019-04-16 11:00:37	https://elpais.com/cultura/2019/04/15/actualidad/1555354154_717447.html
2019-04-16 10:46:57	https://elpais.com/cultura/2019/04/16/actualidad/1555396454_373620.html
2019-04-16 10:11:12	https://elpais.com/cultura/2019/04/15/actualidad/1555351385_404402.html
2019-03-23 23:00:38	https://elpais.com/elpais/2019/03/23/opinion/1553364300_398631.html
2019-03-16 23:00:09	https://elpais.com/elpais/2019/03/16/opinion/1552755194_878279.html
2019-03-16 23:00:00	https://elpais.com/elpais/2019/03/16/opinion/1552748902_089815.html
2019-02-13 11:04:02	https://elpais.com/politica/2019/02/12/actualidad/1549968657_026811.html
2019-02-13 08:41:40	https://elpais.com/politica/2019/02/12/actualidad/1549996308_632843.html
2019-02-01 23:31:58	https://elpais.com/ccaa/2019/01/31/catalunya/1548975109_119530.html

Tabla D-6: Historial de noticias resultante para la temática de El Incendio de la Catedral de Nôtre Dame en el periódico El País

ELECCIONES GENERALES	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,27	0,23	0,25	0,23	0,17	0,29
Titular + Keywords	0,13	0,35	0,35	0,37	0,21	0,38
Titular + Resumen	0,13	0,37	0,29	0,37	0,21	0,38
Titular + Autor	0,23	0,27	0,23	0,21	0,21	0,27
Titular + Cuerpo	0,04	0,29	0,25	0,46	0,15	0,29
Keywords	0,13	0,37	0,31	0,33	0,21	0,37
Keywords + Resumen	0,13	0,35	0,29	0,38	0,21	0,38
Keywords + Autor	0,08	0,27	0,31	0,27	0,17	0,25
Keywords + Cuerpo	0,02	0,27	0,23	0,46	0,17	0,27
Resumen	0,12	0,25	0,1	0,27	0,15	0,27
Resumen + Autor	0,13	0,25	0,17	0,21	0,13	0,25
Resumen + Cuerpo	0,02	0,27	0,25	0,44	0,15	0,27
Autor	0,15	0,12	0,12	0,12	0,15	0,25
Autor + Cuerpo	0,04	0,29	0,27	0,42	0,17	0,29
Cuerpo	0,04	0,27	0,27	0,42	0,17	0,27

Tabla D-7: Resultados para la temática de Las Elecciones Generales del 2019 en el periódico El País

Fecha	Enlace
2019-11-10 00:00:01	https://elpais.com/politica/2019/10/20/actualidad/1571575152_143333.html
2019-10-20 15:29:29	https://elpais.com/politica/2019/10/19/actualidad/1571511156_829840.html
2019-10-20 11:09:36	https://elpais.com/politica/2019/10/18/actualidad/1571419859_044919.html
2019-10-20 07:20:16	https://elpais.com/politica/2019/10/19/actualidad/1571492029_312414.html
2019-10-18 06:52:10	https://elpais.com/tecnologia/2019/10/18/actualidad/1571350153_875274.html
2019-10-17 20:50:09	https://elpais.com/politica/2019/10/17/actualidad/1571330455_082061.html
2019-10-17 12:07:37	https://elpais.com/ccaa/2019/10/17/catalunya/1571295224_497171.html

2019-10-16 23:14:28	https://elpais.com/politica/2019/10/16/actualidad/1571210620_205746.html
2019-10-16 22:16:46	https://elpais.com/deportes/2019/10/16/actualidad/1571222980_611149.html
2019-10-16 17:44:30	https://elpais.com/politica/2019/10/15/actualidad/1571151830_085106.html
2019-10-16 13:22:14	https://elpais.com/politica/2019/10/15/actualidad/1571174717_457801.html
2019-10-16 09:01:40	https://elpais.com/politica/2019/10/15/actualidad/1571162190_293164.html
2019-10-15 19:58:42	https://elpais.com/politica/2019/10/14/actualidad/1571037335_995634.html
2019-10-15 17:41:16	https://elpais.com/politica/2019/10/14/actualidad/1571033446_440448.html
2019-10-13 22:00:29	https://elpais.com/elpais/2019/10/13/opinion/1570980529_717994.html
2019-10-13 17:47:38	https://elpais.com/politica/2019/10/13/actualidad/1570955000_786613.html
2019-10-12 22:00:18	https://elpais.com/elpais/2019/10/12/opinion/1570894707_638767.html
2019-10-10 06:32:42	https://elpais.com/politica/2019/10/09/actualidad/1570633375_966787.html
2019-10-09 19:38:55	https://elpais.com/politica/2019/10/09/actualidad/1570630078_734971.html
2019-10-09 06:17:14	https://elpais.com/politica/2019/10/07/actualidad/1570458030_601164.html
2019-10-08 14:30:18	https://elpais.com/politica/2019/10/08/actualidad/1570521873_087183.html
2019-10-08 13:42:11	https://elpais.com/politica/2019/10/07/actualidad/1570471647_928058.html
2019-10-07 14:14:02	https://elpais.com/politica/2019/10/07/actualidad/1570429352_303738.html
2019-10-07 12:05:38	https://elpais.com/politica/2019/10/06/actualidad/1570388277_919838.html
2019-10-07 08:12:26	https://elpais.com/politica/2019/10/06/actualidad/1570350152_538359.html
2019-10-07 06:59:04	https://elpais.com/politica/2019/10/06/actualidad/1570360893_372899.html
2019-10-06 12:51:45	https://elpais.com/politica/2019/10/05/actualidad/1570291117_117066.html
2019-10-06 10:37:29	https://elpais.com/politica/2019/10/05/actualidad/1570294895_628698.html

2019-10-06 08:50:11	https://elpais.com/politica/2019/10/06/actualidad/1570348639_482720.html
2019-10-05 15:34:28	https://elpais.com/politica/2019/10/04/actualidad/1570217154_572999.html
2019-10-05 09:57:46	https://elpais.com/politica/2019/10/04/actualidad/1570213416_823417.html
2019-10-04 19:53:44	https://elpais.com/politica/2019/10/04/actualidad/1570213436_513464.html
2019-10-04 19:35:42	https://elpais.com/politica/2019/10/04/actualidad/1570181369_310107.html
2019-10-04 13:37:46	https://elpais.com/politica/2019/10/03/actualidad/1570101299_011437.html
2019-10-02 17:56:23	https://elpais.com/politica/2019/10/02/actualidad/1570033740_848051.html
2019-03-22 13:54:44	https://elpais.com/internacional/2019/03/21/actualidad/1553156753_130649.html
2019-03-22 08:36:28	https://elpais.com/internacional/2019/03/21/actualidad/1553205011_253111.html
2019-03-20 20:25:07	https://elpais.com/internacional/2019/03/20/actualidad/1553096561_698877.html
2019-03-19 16:31:12	https://elpais.com/internacional/2019/03/19/actualidad/1552983067_837330.html
2019-02-21 07:34:36	https://elpais.com/politica/2019/02/19/actualidad/1550585655_508262.html
2019-02-15 07:38:25	https://elpais.com/elpais/2019/02/14/opinion/1550169728_099422.html
2019-02-13 16:05:30	https://elpais.com/elpais/2019/02/13/opinion/1550061217_664610.html
2019-02-13 07:49:45	https://elpais.com/politica/2019/02/12/actualidad/1550000816_674719.html
2019-02-12 09:32:17	https://elpais.com/ccaa/2019/02/11/catalunya/1549915314_056223.html
2019-02-12 08:31:45	https://elpais.com/politica/2019/02/11/actualidad/1549913698_215397.html
2019-02-11 17:12:18	https://elpais.com/ccaa/2019/02/10/catalunya/1549824472_635265.html
2019-02-11 10:32:29	https://elpais.com/politica/2019/02/11/actualidad/1549845123_512919.html
2019-02-10 20:48:49	https://elpais.com/politica/2019/02/09/actualidad/1549742828_891549.html
2019-02-06 23:00:13	https://elpais.com/elpais/2019/02/06/opinion/1549476847_703209.html

2019-02-06 16:25:31	https://elpais.com/politica/2019/02/06/actualidad/1549446845_560313.html
2019-02-04 20:22:56	https://elpais.com/ccaa/2019/02/04/catalunya/1549277459_118102.html
2019-02-03 15:37:01	https://elpais.com/politica/2019/02/02/actualidad/1549136134_435356.html
2019-02-01 20:10:37	https://elpais.com/politica/2019/02/01/actualidad/1549016626_676859.html

Tabla D-8: Historial de noticias resultante para la temática de Las Elecciones Generales del 2019 en el periódico El País

EXHUMACIÓN FRANCO	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,67	0,67	0,67	0,67	0,33	0,73
Titular + Keywords	0,67	0,73	0,47	0,8	0,67	0,73
Titular + Resumen	0,6	0,8	0,6	0,8	0,4	0,8
Titular + Autor	0,67	0,73	0,67	0,73	0,4	0,73
Titular + Cuerpo	0,4	0,47	0,33	0,67	0,33	0,47
Keywords	0,67	0,73	0,67	0,67	0,67	0,73
Keywords + Resumen	0,4	0,73	0,73	0,73	0,53	0,73
Keywords + Autor	0,73	0,73	0,67	0,73	0,6	0,73
Keywords + Cuerpo	0,4	0,47	0,4	0,6	0,33	0,47
Resumen	0,33	0,73	0,73	0,73	0,47	0,73
Resumen + Autor	0,33	0,67	0,6	0,6	0,4	0,67
Resumen + Cuerpo	0,4	0,47	0,33	0,6	0,33	0,47
Autor	0,47	0,33	0,33	0,33	0,47	0,6
Autor + Cuerpo	0,4	0,53	0,33	0,6	0,33	0,53
Cuerpo	0,4	0,47	0,33	0,6	0,33	0,47

Tabla D-9: Resultados para la temática de La Exhumación del Dictador Francisco Franco en el periódico 20 Minutos

Fecha	Enlace
2019-10-25 21:51:42	https://www.20minutos.es/noticia/4030716/0/sanchez-confiesa-exhumacion-franco-provoco-enorme-orgullo-emocion/
2019-10-25 18:31:44	https://www.20minutos.es/noticia/4030651/0/significado-tiene-bandera-cubria-feretro-francisco-franco/
2019-10-25 14:54:19	https://www.20minutos.es/noticia/4030271/0/queipo-llano-golpista-cuyos-restos-seran-proximamente-exhumados-basilica-macarena-sevilla/
2019-10-25 13:35:07	https://www.20minutos.es/videos/nacional/4030102-tension-a-la-salida-de-la-familia-franco-de-mingorrubio/
2019-10-25 13:26:44	https://www.20minutos.es/noticia/4030056/0/gobierno-analiza-viva-franco/
2019-10-25 12:34:55	https://www.20minutos.es/noticia/4029928/0/gobierno-valle-caidos-recordar-victimas/
2019-10-24 21:10:02	https://www.20minutos.es/videos/economia/4029147-ultraderechistas-homenajean-al-dictador-francisco-franco/

2019-10-23 17:31:28	https://www.20minutos.es/videos/economia/4027100-el-gobierno-instalara-un-escaner-para-que-no-se-grabe-la-exhumacion/
2019-10-23 16:46:59	https://www.20minutos.es/noticia/4027006/0/abascal-muertos-tutankamon/
2019-10-23 16:38:04	https://www.20minutos.es/noticia/4026908/0/iglesias-colau-reunen-barcelona/
2019-10-20 06:39:13	https://www.20minutos.es/noticia/3806619/0/helicoptero-forense-sin-honores-gobierno-ultima-exhumacion-franco-valle-caidos/
2019-10-15 18:26:36	https://www.20minutos.es/noticia/3802330/0/supremo-estudiara-otro-recurso-fundacion-nacional-francisco-franco-contra-exhumacion/
2019-10-13 11:25:54	https://www.20minutos.es/noticia/3798993/0/gobierno-contrata-funeraria-exhumacion-franco/
2019-10-12 16:10:43	https://www.20minutos.es/noticia/3798693/0/pedro-sanchez-imviable-perder-elecciones-generales/
2019-10-12 13:19:20	https://www.20minutos.es/noticia/3798562/0/guardia-civil-impide-entrada-valle-caidos/
2019-10-02 12:13:35	https://www.20minutos.es/noticia/3786030/0/relator-155-ley-seguridad-nacional-giro-sanchez-cataluna-10n/

Tabla D-10: Historial de noticias resultante para la temática de La Exhumación del Dictador Francisco Franco en el periódico 20 Minutos

KOBE BRYANT	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,1	0,85	0,85	0,7	0	0,84
Titular + Keywords	0,5	1	0,95	1	0	1
Titular + Resumen	0,15	0,8	0,85	0,8	0,05	0,79
Titular + Autor	0,1	0,8	0,6	0,8	0	0,79
Titular + Cuerpo	0,3	0,7	0,8	0,95	0	0,68
Keywords	0,8	0,9	0,95	1	0	0,95
Keywords + Resumen	0,65	0,9	0,85	0,95	0	0,95
Keywords + Autor	0,85	0,95	0,75	1	0	0,95
Keywords + Cuerpo	0,25	0,8	0,85	0,95	0	0,79
Resumen	0	0	0	0	0,21	0,21
Resumen + Autor	0,05	0,15	0,05	0,15	0	0,21
Resumen + Cuerpo	0,3	0,65	0,8	0,9	0	0,63
Autor	0,1	0,1	0,3	0,1	0	0,11
Autor + Cuerpo	0,25	0,7	0,85	0,9	0	0,68
Cuerpo	0,25	0,7	0,85	0,9	0	0,68

Tabla D-11: Resultados para la temática de La Muerte de Kobe Bryant en el periódico El Mundo

Fecha	Enlace
2020-02-07T21:18:51Z	https://www.elmundo.es/deportes/baloncesto/nba/2020/02/07/5e3dd42c21efa035088b459f.html
2020-02-02 09:56:12	https://www.elmundo.es/deportes/baloncesto/nba/2020/02/02/5e369c8d21efa01c0a8b4660.html
2020-02-01 08:58:24	https://www.elmundo.es/deportes/baloncesto/nba/2020/02/01/5e353da021efa0f7248b45f5.html
2020-01-31 09:15:02	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/31/5e33f004fc6c8336478b45a1.html
2020-01-31 08:36:12	https://www.elmundo.es/deportes/baloncesto/2020/01/31/5e33e4c9fdddff1a3a8b45a3.html
2020-01-30 22:01:09	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/30/5e335216fdddff641f8b459e.html
2020-01-30 06:57:18	https://www.elmundo.es/deportes/baloncesto/2020/01/30/5e327dbcfdddff168c8b45f4.html

2020-01-29 09:11:09	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/29/5e314c1ffddddff01428b45b0.html
2020-01-29 08:00:00	https://www.elmundo.es/deportes/baloncesto/2020/01/29/5e313a50fdddff21638b4594.html
2020-01-28 22:49:57	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/28/5e308def21efa098578b469e.html
2020-01-28 12:20:45	https://www.elmundo.es/deportes/tenis/open-de-australia/2020/01/28/5e30270ffdddff73428b4658.html
2020-01-28 09:48:51	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/28/5e30037321efa098578b460c.html
2020-01-28 01:21:42	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/28/5e2f214f21efa00e608b45e1.html
2020-01-27 20:58:06	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/27/5e2f4ecc21efa01f178b45d2.html
2020-01-27 16:58:23	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/27/5e2f142b21efa0b3668b45fd.html
2020-01-27 16:11:50	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/27/5e2efc0ffdddff43a68b462a.html
2020-01-27 10:12:48	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/27/5e2eb791fc6c83ee3c8b4629.html
2020-01-27 09:31:19	https://www.elmundo.es/deportes/baloncesto/2020/01/27/5e2ea41421efa0273a8b4575.html
2020-01-27 07:18:10	https://www.elmundo.es/deportes/baloncesto/nba/2020/01/27/5e2e8e33fc6c83cb278b459d.html
2020-01-27 00:14:58	https://www.elmundo.es/deportes/baloncesto/2020/01/27/5e2e251321efa05c6f8b46c3.html
2020-01-26 21:16:08	https://www.elmundo.es/deportes/baloncesto/2020/01/26/5e2e018321efa0b6338b4693.html

Tabla D-12: Historial de noticias resultante para la temática de La Muerte de Kobe Bryant en el periódico El Mundo

DÍA INTERNACIONAL DE LA MUJER	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,41	0,47	0,41	0,41	0,12	0,47
Titular + Keywords	0,82	0,76	0,59	0,41	0,24	0,82
Titular + Resumen	0,47	0,24	0,29	0,18	0,24	0,24
Titular + Autor	0,41	0,47	0,41	0,41	0,12	0,47
Titular + Cuerpo	0,24	0,53	0,47	0,65	0,06	0,53
Keywords	0,82	0,59	0,71	0,53	0,41	0,59
Keywords + Resumen	0,82	0,65	0,59	0,24	0,18	0,65
Keywords + Autor	0,82	0,59	0,71	0,41	0,47	0,59
Keywords + Cuerpo	0,35	0,59	0,47	0,76	0,12	0,59
Resumen	0,24	0,18	0,29	0,18	0,24	0,18
Resumen + Autor	0,29	0,12	0,29	0,18	0,12	0,18
Resumen + Cuerpo	0,29	0,53	0,53	0,71	0,06	0,53
Autor	0	0	0	0	0,06	0,41
Autor + Cuerpo	0,29	0,53	0,41	0,65	0,06	0,53
Cuerpo	0,29	0,53	0,41	0,65	0,06	0,53

Tabla D-13: Resultados para la temática de El Día Internacional de la Mujer del 2020 en el periódico 20 Minutos

Fecha	Enlace
2020-03-08 23:04:39	https://www.20minutos.es/noticia/4180255/0/la-revuelta-feminista-toma-espana-pero-con-menor-asistencia-que-en-los-dos-ultimos-anos/
2020-03-08 18:26:17	https://www.20minutos.es/videos/nacional/4180181-ciudadanos-abandona-marcha-8m-empujones-tension/
2020-03-08 18:13:11	https://www.20minutos.es/noticia/4180163/0/ciudadanos-obligado-irse-marcha-8m-increpados-empujones-tension/
2020-03-08 17:01:00	https://www.20minutos.es/noticia/4180129/0/asi-viven-manifestaciones-8-marzo-espana/
2020-03-08 12:05:16	https://www.20minutos.es/noticia/4179924/0/sanidad-pide-personas-sintomas-no-acudan-manifestaciones-8m/
2020-03-08 11:16:33	https://www.20minutos.es/noticia/4179867/0/dia-internacional-mujer-medidas-prevencion-manifestaciones-8-marzo-coronavirus/
2020-03-08 10:48:25	https://www.20minutos.es/noticia/4179839/0/dia-internacional-mujer-2020-significa-feminismo-feminista/

2020-03-08 10:09:00	https://www.20minutos.es/noticia/4178413/0/dia-internacional-mujer-color-morado-lucha-feminista/
2020-03-08 08:10:31	https://www.20minutos.es/noticia/4171345/0/feminismo-huelga-feminista-8-marzo-manifestaciones-directo/
2020-03-08 08:01:02	https://www.20minutos.es/noticia/4174712/0/feminismo-nomo-generacion-mujeres/
2020-03-07 17:55:15	https://www.20minutos.es/noticia/4177707/0/denunciar-violencia-genero-telefonos-servicios-disposicion-mujeres/
2020-03-07 17:10:00	https://www.20minutos.es/noticia/4173655/0/origen-manifestaciones-8m-paro-internacional-mujeres/
2020-03-07 08:23:24	https://www.20minutos.es/noticia/4175293/0/feminismo-madrid-horario-recorrido-manifestacion-8m/
2020-03-07 08:21:43	https://www.20minutos.es/noticia/4172105/0/feminismo-gisella-perl-auschwitz-embarazos/
2020-03-06 19:19:28	https://www.20minutos.es/noticia/4178120/0/diferencias-feminismo-radical-liberal-materias-sociales/
2020-03-05 15:50:43	https://www.20minutos.es/noticia/4174770/0/francia-alemania-portugal-manifestaciones-8-marzo-europa/
2020-03-05 07:52:47	https://www.20minutos.es/noticia/4174425/0/dia-internacional-mujer-2020-feminismo-horarios-recorridos-8m-espana/
2020-02-03 12:41:48	https://www.20minutos.es/noticia/4139525/0/no-soy-un-virus-asiaticos-denuncian-racismo-coronavirus/

Tabla D-14: Historial de noticias resultante para la temática de El Día Internacional de la Mujer del 2020 en el periódico 20 Minutos

CORONAVIRUS	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,57	0,76	0,73	0,76	0,46	0,78
Titular + Keywords	0,71	0,74	0,65	0,96	0,41	0,75
Titular + Resumen	0,51	0,75	0,72	0,7	0,46	0,76
Titular + Autor	0,5	0,73	0,72	0,73	0,41	0,74
Titular + Cuerpo	0,66	0,75	0,58	0,71	0,47	0,75
Keywords	0,66	0,76	0,62	0,92	0,34	0,82
Keywords + Resumen	0,68	0,71	0,59	0,87	0,49	0,71
Keywords + Autor	0,67	0,74	0,58	0,9	0,33	0,56
Keywords + Cuerpo	0,67	0,76	0,58	0,75	0,48	0,76
Resumen	0,47	0,41	0,28	0,41	0,43	0,45
Resumen + Autor	0,44	0,39	0,16	0,4	0,39	0,42
Resumen + Cuerpo	0,66	0,75	0,58	0,68	0,46	0,75
Autor	0,46	0,22	0,33	0,22	0,46	0,77
Autor + Cuerpo	0,66	0,74	0,58	0,69	0,48	0,74
Cuerpo	0,66	0,75	0,58	0,69	0,48	0,75

Tabla D-15: Resultados para la temática de La COVID-19 en el periódico El Mundo

Fecha	Enlace
2020-03-17T01:10:42Z	https://www.elmundo.es/espana/2020/03/17/5e6fd1fcfdddff5f1c8b45b3.html
2020-03-17 01:10:42	https://www.elmundo.es/economia/macroeconomia/2020/03/17/5e6f948121efa05e088b462a.html
2020-03-17 01:10:42	https://www.elmundo.es/economia/2020/03/17/5e6fe422fdddff3db98b45b8.html
2020-03-17 01:10:42	https://www.elmundo.es/economia/macroeconomia/2020/03/17/5e6fe948121efa0e9398b4617.html
2020-03-16 21:19:26	https://www.elmundo.es/economia/macroeconomia/2020/03/16/5e6fed4ffdddffe3698b4590.html
2020-03-16 21:05:52	https://www.elmundo.es/internacional/2020/03/16/5e6fea0121efa0302a8b45e7.html
2020-03-16 20:40:32	https://www.elmundo.es/economia/macroeconomia/2020/03/16/5e6fe41bfc6c83af1e8b4612.html

2020-03-16 19:41:15	https://www.elmundo.es/cultura/cine/2020/03/16/5e6fd57021efa0b4368b45d5.html
2020-03-16 19:26:31	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/16/5e6fd1f321efa034388b45c9.html
2020-03-16 19:20:30	https://www.elmundo.es/motor/2020/03/16/5e6fd16efc6c830a308b4611.html
2020-03-16 18:30:48	https://www.elmundo.es/internacional/2020/03/16/5e6fc59bfc6c83bc758b45f8.html
2020-03-16 17:59:16	https://www.elmundo.es/ciencia-y-salud/ciencia/2020/03/16/5e6fac08fdddff84968b45a5.html
2020-03-16 17:45:49	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/16/5e6fbad421efa0075b8b4655.html
2020-03-16 16:11:35	https://www.elmundo.es/internacional/2020/03/16/5e6fa4a621efa0c80c8b45b3.html
2020-03-16 16:02:33	https://www.elmundo.es/deportes/baloncesto/2020/03/16/5e6fa2ae21efa090418b45cf.html
2020-03-16 14:10:45	https://www.elmundo.es/cataluna/2020/03/16/5e6f88d6fc6c834e438b45d4.html
2020-03-16 13:17:37	https://www.elmundo.es/television/2020/03/16/5e6f7c59fc6c83af1e8b45cf.html
2020-03-16 12:42:55	https://www.elmundo.es/economia/2020/03/16/5e6f51d021efa0fd268b458c.html
2020-03-16 12:35:04	https://www.elmundo.es/madrid/2020/03/16/5e6f711bfdddff0a768b456e.html
2020-03-16 12:17:32	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/16/5e6f6e14fc6c838a7c8b45b9.html
2020-03-16 11:49:54	https://www.elmundo.es/madrid/2020/03/16/5e6f672efdddff85048b464e.html
2020-03-16 11:38:30	https://www.elmundo.es/andalucia/2020/03/16/5e6f6509fc6c83cc3a8b45a5.html
2020-03-16 11:15:15	https://www.elmundo.es/internacional/2020/03/16/5e6f5fb3fc6c83fc478b459a.html
2020-03-16 11:02:48	https://www.elmundo.es/deportes/mas-deporte/2020/03/16/5e6f5c74fdddff93948b45f6.html
2020-03-16 10:24:37	https://www.elmundo.es/madrid/2020/03/16/5e6f5264fc6c83af1e8b45b4.html
2020-03-16 10:02:50	https://www.elmundo.es/andalucia/2020/03/16/5e6f4eb8fdddff93948b45e5.html
2020-03-16 10:01:47	https://www.elmundo.es/cultura/2020/03/16/5e6b76cc21efa07c058b4572.html
2020-03-16 10:00:21	https://www.elmundo.es/economia/macroeconomia/2020/03/16/5e6f4de2fc6c838a7c8b45a3.html

2020-03-16 09:54:37	https://www.elmundo.es/espana/2020/03/16/5e6f49bafdddffff038b45ea.html
2020-03-16 09:51:26	https://www.elmundo.es/madrid/2020/03/16/5e6f4b2cfc6c838a7c8b459a.html
2020-03-16 08:45:36	https://www.elmundo.es/deportes/baloncesto/nba/2020/03/16/5e6f3c80fdddff603c8b458a.html
2020-03-16 08:38:21	https://www.elmundo.es/espana/2020/03/16/5e6f2c2821efa003678b4574.html
2020-03-16 08:18:05	https://www.elmundo.es/deportes/futbol/2020/03/16/5e6f362ffdddffc4668b4657.html
2020-03-16 07:30:16	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/16/5e6f28defdddffd8058b4630.html
2020-03-16 01:21:43	https://www.elmundo.es/viajes/el-baul/2020/03/16/5e6b6d0bfc6c83580d8b4595.html
2020-03-16 01:21:15	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/16/5e6a337f21efa0aa138b45fd.html
2020-03-16 01:21:12	https://www.elmundo.es/deportes/baloncesto/2020/03/16/5e6e5b3321efa0be3d8b4793.html
2020-03-16 01:20:52	https://www.elmundo.es/economia/2020/03/16/5e6e95c6fc6c8366528b4573.html
2020-03-16 00:48:31	https://www.elmundo.es/madrid/2020/03/16/5e6e691bfc6c83306d8b45db.html
2020-03-15 23:31:53	https://www.elmundo.es/internacional/2020/03/16/5e6ebad7fc6c83bc758b458c.html
2020-03-15 21:28:41	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/15/5e6e9d7421efa0ad1c8b47e2.html
2020-03-15 20:53:36	https://www.elmundo.es/espana/2020/03/15/5e6e7df5fc6c830a308b457f.html
2020-03-15 20:49:10	https://www.elmundo.es/espana/2020/03/15/5e6e9481fdddffb54d8b4615.html
2020-03-15 20:27:13	https://www.elmundo.es/internacional/2020/03/15/5e6e8ceb21efa090418b4580.html
2020-03-15 19:29:22	https://www.elmundo.es/madrid/2020/03/15/5e6e7ea5fc6c8391578b456e.html
2020-03-15 19:09:49	https://www.elmundo.es/andalucia/2020/03/15/5e6e7d6bfdddfffe8d8b457b.html
2020-03-15 18:39:51	https://www.elmundo.es/baleares/2020/03/15/5e6e7677fdddff38548b45f1.html
2020-03-15 17:56:27	https://www.elmundo.es/economia/macroeconomia/2020/03/15/5e6e6c3621efa0fc1d8b4756.html
2020-03-15 17:56:10	https://www.elmundo.es/espana/2020/03/15/5e6e6bfcfdddffd2058b45fb.html

2020-03-15 17:23:51	https://www.elmundo.es/baleares/2020/03/15/5e6e648bfdddff6aa98b45de.html
2020-03-15 17:12:26	https://www.elmundo.es/espana/2020/03/15/5e6e5abb21efa0075b8b45be.html
2020-03-15 17:11:14	https://www.elmundo.es/cataluna/2020/03/15/5e6e61a0fdddffba6b8b4624.html
2020-03-15 16:35:28	https://www.elmundo.es/cultura/2020/03/15/5e6e565afc6c833c4c8b471c.html
2020-03-15 16:30:09	https://www.elmundo.es/andalucia/2020/03/15/5e6e580121efa0c83d8b46c7.html
2020-03-15 16:17:41	https://www.elmundo.es/madrid/2020/03/15/5e6e5281fc6c83c90e8b4760.html
2020-03-15 15:01:27	https://www.elmundo.es/espana/2020/03/15/5e6e3d60fdddffb54d8b45e6.html
2020-03-15 14:12:54	https://www.elmundo.es/andalucia/2020/03/15/5e6e37d621efa0b07e8b4586.html
2020-03-15 12:46:11	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/15/5e6e200dfc6c83983d8b4749.html
2020-03-15 12:18:54	https://www.elmundo.es/espana/2020/03/15/5e6e0fb221efa0e53d8b47a8.html
2020-03-15 12:12:27	https://www.elmundo.es/deportes/futbol/2020/03/15/5e6e1b99fc6c830c6c8b4734.html
2020-03-15 12:10:41	https://www.elmundo.es/internacional/2020/03/15/5e6e156ffdddff6d798b45ae.html
2020-03-15 11:48:22	https://www.elmundo.es/deportes/futbol/2020/03/15/5e6e15e5fdddff6aa98b45be.html
2020-03-15 10:56:35	https://www.elmundo.es/espana/2020/03/15/5e6e09d221efa07c058b47e8.html
2020-03-15 10:50:49	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/15/5e6d35c921efa0c1348b46ba.html
2020-03-15 10:43:23	https://www.elmundo.es/espana/2020/03/15/5e6e06a5fdddff51558b466a.html
2020-03-15 09:51:25	https://www.elmundo.es/espana/2020/03/15/5e6dfa8a21efa07c248b470c.html
2020-03-15 07:52:46	https://www.elmundo.es/internacional/2020/03/15/5e6dde2521efa0ab3b8b47eb.html
2020-03-15 07:33:40	https://www.elmundo.es/espana/2020/03/15/5e6dda23fc6c83580d8b4699.html
2020-03-15 07:16:53	https://www.elmundo.es/comunidad-valenciana/2020/03/15/5e6d24c6fc6c83c90e8b4719.html
2020-03-15 01:56:12	https://www.elmundo.es/cultura/cine/2020/03/15/5e6d8147fc6c83c90e8b473b.html

2020-03-15 01:55:14	https://www.elmundo.es/cultura/musica/2020/03/15/5e6ce45a21efa015078b456f.html
2020-03-15 01:40:06	https://www.elmundo.es/comunidad-valenciana/castellon/2020/03/15/5e6d4f3b21efa0c7088b4730.html
2020-03-15 01:25:07	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/15/5e6d3a82fc6c8351618b46e5.html
2020-03-15 01:02:08	https://www.elmundo.es/madrid/2020/03/15/5e6d27fafdddff7d068b45be.html
2020-03-14 22:59:23	https://www.elmundo.es/espana/2020/03/14/5e6d61bbfdddff38548b459d.html
2020-03-14 21:16:17	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/14/5e6d4991fdddff6d798b458a.html
2020-03-14 21:13:52	https://www.elmundo.es/espana/2020/03/14/5e6d48c421efa0c1348b46d1.html
2020-03-14 20:47:29	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/14/5e6d3f4e21efa0fb148b47e6.html
2020-03-14 19:17:01	https://www.elmundo.es/vida-sana/bienestar/2020/03/14/5e6b9a53fdddff2d8a8b4619.html
2020-03-14 18:45:32	https://www.elmundo.es/f5/comparte/2020/03/14/5e6d212921efa0903d8b458e.html
2020-03-14 17:05:19	https://www.elmundo.es/internacional/2020/03/14/5e6d0dc2fc6c83570d8b46ad.html
2020-03-14 16:42:33	https://www.elmundo.es/espana/2020/03/14/5e6d07f3fdddff01148b4573.html
2020-03-14 16:39:10	https://www.elmundo.es/espana/2020/03/14/5e6d089dfdddff7aa08b4657.html
2020-03-14 16:20:07	https://www.elmundo.es/economia/empresas/2020/03/14/5e6d03d7fdddff13918b465f.html
2020-03-14 15:11:26	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/14/5e6942b221efa0ff018b45cc.html
2020-03-14 14:57:39	https://www.elmundo.es/internacional/2020/03/14/5e6cf0d3fc6c83586d8b46f7.html
2020-03-14 14:38:50	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/14/5e6cec20fdddffc1798b459a.html
2020-03-14 14:04:19	https://www.elmundo.es/pais-vasco/2020/03/14/5e6ce452fc6c83586d8b46e8.html
2020-03-14 12:55:12	https://www.elmundo.es/espana/2020/03/14/5e6cd044fdddff58058b4595.html
2020-03-14 12:48:58	https://www.elmundo.es/espana/2020/03/14/5e6cb1fefdddff13918b463f.html
2020-03-14 12:47:49	https://www.elmundo.es/andalucia/2020/03/14/5e6cd26421efa0106f8b46c6.html

2020-03-14 11:54:52	https://www.elmundo.es/internacional/2020/03/14/5e6cc5f921efa07f6b8b470a.html
2020-03-14 10:54:42	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/14/5e6cb3ab21efa0ab3b8b47a0.html
2020-03-14 10:20:32	https://www.elmundo.es/madrid/2020/03/14/5e6cafb121efa0ad1c8b4728.html
2020-03-14 10:11:59	https://www.elmundo.es/espana/2020/03/14/5e6cacacfdddffba6b8b45a1.html
2020-03-14 01:11:14	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/14/5e6b6c8421efa0e07e8b45fd.html
2020-03-14 00:58:34	https://www.elmundo.es/economia/macroeconomia/2020/03/14/5e6bcc2e21efa0fc1d8b467a.html
2020-03-14 00:58:11	https://www.elmundo.es/espana/2020/03/14/5e6be5d3fc6c83cf598b46ac.html
2020-03-14 00:10:35	https://www.elmundo.es/cultura/musica/2020/03/14/5e6bc2fafc6c8319268b46ec.html
2020-03-13 20:09:08	https://www.elmundo.es/economia/empresas/2020/03/13/5e6be7ae21efa0e53d8b4717.html
2020-03-13 19:34:46	https://www.elmundo.es/internacional/2020/03/13/5e6bc586fdddffa6598b4672.html
2020-03-13 19:14:45	https://www.elmundo.es/espana/2020/03/13/5e6bda5bfdddff84b08b4576.html
2020-03-13 18:53:06	https://www.elmundo.es/cataluna/2020/03/13/5e6bd64dfc6c83983d8b46d5.html
2020-03-13 18:09:11	https://www.elmundo.es/internacional/2020/03/13/5e6bcb7afc6c83d3118b4639.html
2020-03-13 17:35:54	https://www.elmundo.es/tecnologia/2020/03/13/5e6bc1a5fdddffa6598b4668.html
2020-03-13 17:32:04	https://www.elmundo.es/motor/2020/03/13/5e6bc372fdddff7aa08b460a.html
2020-03-13 17:30:35	https://www.elmundo.es/espana/2020/03/13/5e6bc327fdddffcc3c8b4692.html
2020-03-13 17:09:39	https://www.elmundo.es/deportes/ciclismo/2020/03/13/5e6bbe03fc6c838c3d8b4587.html
2020-03-13 16:15:32	https://www.elmundo.es/cultura/2020/03/13/5e6bb19421efa0e53d8b46d2.html
2020-03-13 16:03:18	https://www.elmundo.es/vida-sana/familia-y-co/2020/03/13/5e6ba7f5fdddffb1168b46a5.html
2020-03-13 16:02:15	https://www.elmundo.es/internacional/2020/03/13/5e6bae7621efa00e788b4745.html
2020-03-13 15:59:02	https://www.elmundo.es/economia/macroeconomia/2020/03/13/5e6badb221efa0ad1c8b46d6.html

2020-03-13 15:54:50	https://www.elmundo.es/vida-sana/cuerpo/2020/03/13/5e6b93a3fdddff1f4c8b46a5.html
2020-03-13 15:34:48	https://www.elmundo.es/espana/2020/03/13/5e6ba657fdddff22158b4695.html
2020-03-13 14:58:38	https://www.elmundo.es/espana/2020/03/13/5e6b9f8d21efa0c1348b460c.html
2020-03-13 14:27:54	https://www.elmundo.es/cataluna/2020/03/13/5e6b975a21efa07c248b4605.html
2020-03-13 13:43:59	https://www.elmundo.es/espana/2020/03/13/5e6b8e0521efa0fb148b45b3.html
2020-03-13 13:20:24	https://www.elmundo.es/espana/2020/03/13/5e6b74e5fc6c83330c8b459e.html
2020-03-13 13:13:51	https://www.elmundo.es/espana/2020/03/13/5e6b844e21efa0dd258b45a5.html
2020-03-13 12:58:39	https://www.elmundo.es/cultura/2020/03/13/5e6b830421efa0fa068b45bc.html
2020-03-13 12:03:19	https://www.elmundo.es/madrid/2020/03/13/5e6b705c21efa0252e8b459f.html
2020-03-13 11:19:01	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/13/5e6b6879fdddff10968b4645.html
2020-03-13 10:22:49	https://www.elmundo.es/espana/2020/03/13/5e6b4e2921efa06a0b8b4671.html
2020-03-13 10:00:25	https://www.elmundo.es/pais-vasco/2020/03/13/5e6b59a621efa0b34a8b45b7.html
2020-03-13 09:31:22	https://www.elmundo.es/espana/2020/03/13/5e6b50af21efa02a3a8b45dd.html
2020-03-13 09:24:34	https://www.elmundo.es/andalucia/sevilla/2020/03/13/5e6b4dd521efa08a628b4660.html
2020-03-13 09:08:32	https://www.elmundo.es/madrid/2020/03/13/5e6b4d8121efa06a0b8b4669.html
2020-03-13 08:24:13	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/13/5e6b4230fdddff878c8b4624.html
2020-03-13 08:12:51	https://www.elmundo.es/economia/macroeconomia/2020/03/13/5e6b4064fc6c83d2118b4593.html
2020-03-13 07:50:21	https://www.elmundo.es/cataluna/2020/03/13/5e6b39d521efa072518b467b.html
2020-03-13 01:10:46	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/13/5e6aa5f9fc6c83c90e8b456e.html
2020-03-13 01:10:45	https://www.elmundo.es/f5/comparte/2020/03/13/5e6a20fffc6c83796c8b4699.html
2020-03-13 01:10:27	https://www.elmundo.es/internacional/2020/03/13/5e6ad0ff21efa0941a8b463b.html

2020-03-13 01:04:11	https://www.elmundo.es/cultura/cine/2020/03/13/5e6a7a4cfc6c83703b8b4613.html
2020-03-13 00:42:49	https://www.elmundo.es/opinion/2020/03/13/5e6a8d3cfdddff7ba08b460b.html
2020-03-12 20:33:39	https://www.elmundo.es/cataluna/2020/03/12/5e6a9c9121efa0941a8b462b.html
2020-03-12 20:25:14	https://www.elmundo.es/internacional/2020/03/12/5e6a947ffdddff7aa08b45ae.html
2020-03-12 19:48:02	https://www.elmundo.es/madrid/2020/03/12/5e6a7ab5fdddff2d8a8b45cf.html
2020-03-12 19:04:11	https://www.elmundo.es/cultura/teatro/2020/03/12/5e6a870bfc6c83b8058b4813.html
2020-03-12 18:10:36	https://www.elmundo.es/motor/2020/03/12/5e6a7b0cfc6c8364518b479a.html
2020-03-12 17:17:08	https://www.elmundo.es/madrid/2020/03/12/5e6a6ddffddddffe4798b45d6.html
2020-03-12 16:58:22	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/12/5e6a652afc6c83d3538b4722.html
2020-03-12 16:07:03	https://www.elmundo.es/deportes/futbol/champions-league/2020/03/12/5e6a5e17fdddff2b418b4718.html
2020-03-12 16:04:49	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/12/5e6a29d8fc6c83b8438b46b2.html
2020-03-12 16:04:05	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/12/5e6a5b4121efa063218b4614.html
2020-03-12 15:54:53	https://www.elmundo.es/deportes/tenis/2020/03/12/5e6a5b3afdddff698e8b45fd.html
2020-03-12 15:40:19	https://www.elmundo.es/economia/macroeconomia/2020/03/12/5e6a57b621efa06a0b8b4617.html
2020-03-12 15:20:56	https://www.elmundo.es/espana/2020/03/12/5e6a4758fdddff878c8b45e8.html
2020-03-12 14:44:08	https://www.elmundo.es/cultura/2020/03/12/5e6a49a4fdddffa2908b45f4.html
2020-03-12 14:19:17	https://www.elmundo.es/madrid/2020/03/12/5e6a445efdddff7aa08b4575.html
2020-03-12 13:38:36	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/12/5e6a3b4cfc6c83ac098b47ce.html
2020-03-12 12:29:39	https://www.elmundo.es/economia/empresas/2020/03/12/5e6a28d721efa063218b45f8.html
2020-03-12 12:26:18	https://www.elmundo.es/internacional/2020/03/12/5e6a2a4bfdddff10968b45c3.html
2020-03-12 12:14:11	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/12/5e6a23c721efa0c0628b4629.html

2020-03-12 12:12:52	https://www.elmundo.es/madrid/2020/03/12/5e6a1227fc6c83d3538b46f1.html
2020-03-12 12:07:18	https://www.elmundo.es/espana/2020/03/12/5e6a25e6fdddff10968b45ba.html
2020-03-12 11:44:35	https://www.elmundo.es/espana/2020/03/12/5e6a2011fdddff2c1e8b4574.html
2020-03-12 11:29:10	https://www.elmundo.es/cultura/2020/03/12/5e6a19e3fdddffac5a8b4755.html
2020-03-12 11:17:22	https://www.elmundo.es/espana/2020/03/12/5e6a1a2bfdddff18b88b4700.html
2020-03-12 11:06:29	https://www.elmundo.es/deportes/futbol/primera-division/2020/03/12/5e6a1799fc6c83ac098b47b6.html
2020-03-12 10:41:20	https://www.elmundo.es/deportes/baloncesto/liga-endesa/2020/03/12/5e6a11be21efa0210c8b4731.html
2020-03-12 10:21:59	https://www.elmundo.es/internacional/2020/03/12/5e6a0a77fdddff3e968b45c5.html
2020-03-12 10:05:54	https://www.elmundo.es/espana/2020/03/12/5e6a095a21efa06f668b45f3.html
2020-03-12 09:03:55	https://www.elmundo.es/espana/2020/03/12/5e69fae321efa054508b45d4.html
2020-03-12 08:47:00	https://www.elmundo.es/madrid/2020/03/12/5e69f5be21efa01f108b4607.html
2020-03-12 08:44:54	https://www.elmundo.es/deportes/futbol/2020/03/12/5e691e31fc6c8345568b4717.html
2020-03-12 08:25:20	https://www.elmundo.es/internacional/2020/03/12/5e69f1dffdddffdc218b4669.html
2020-03-12 08:22:37	https://www.elmundo.es/economia/macroeconomia/2020/03/12/5e69f13e21efa0fe658b45cb.html
2020-03-12 08:10:07	https://www.elmundo.es/madrid/2020/03/12/5e69ed6921efa0941a8b45be.html
2020-03-12 01:07:10	https://www.elmundo.es/cultura/literatura/2020/03/12/5e68defc21efa0210c8b46a9.html
2020-03-12 01:07:01	https://www.elmundo.es/espana/2020/03/12/5e694cca21efa01c538b45c0.html
2020-03-11 19:51:52	https://www.elmundo.es/opinion/2020/03/11/5e693fea21efa0332d8b45c9.html
2020-03-11 19:42:59	https://www.elmundo.es/internacional/2020/03/11/5e693f08fc6c83a31b8b4750.html
2020-03-11 19:42:30	https://www.elmundo.es/madrid/2020/03/11/5e69342421efa06a0b8b45a8.html
2020-03-11 17:36:15	https://www.elmundo.es/internacional/2020/03/11/5e691f0afc6c8387298b456f.html

2020-03-11 16:55:14	https://www.elmundo.es/internacional/2020/03/11/5e690c6721efa08c488b4573.html
2020-03-11 14:52:05	https://www.elmundo.es/comunidad-valenciana/2020/03/11/5e68fb01fdddff2b418b4671.html
2020-03-11 14:44:19	https://www.elmundo.es/economia/macroeconomia/2020/03/11/5e68f93121efa0f1408b4776.html
2020-03-11 14:14:20	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/11/5e68f1a121efa058068b465b.html
2020-03-11 13:45:42	https://www.elmundo.es/madrid/2020/03/11/5e68e78721efa0950d8b4758.html
2020-03-11 13:28:15	https://www.elmundo.es/deportes/futbol/2020/03/11/5e68e75dfdddff02678b472f.html
2020-03-11 12:52:27	https://www.elmundo.es/espana/2020/03/11/5e68defcfc6c8329248b4709.html
2020-03-11 12:51:51	https://www.elmundo.es/vida-sana/cuerpo/2020/03/11/5e68dbbf21efa058068b4648.html
2020-03-11 12:48:01	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/11/5e68dcf4fc6c83a31b8b46d7.html
2020-03-11 12:31:11	https://www.elmundo.es/deportes/futbol/2020/03/11/5e68d9fcfdddff271d8b479b.html
2020-03-11 10:46:05	https://www.elmundo.es/deportes/formula-1/2020/03/11/5e68c15dfdddfff7168b4698.html
2020-03-11 10:39:46	https://www.elmundo.es/espana/2020/03/11/5e68bfde21efa0511f8b464b.html
2020-03-11 10:08:08	https://www.elmundo.es/cultura/musica/2020/03/11/5e68b679fc6c8328198b45b0.html
2020-03-11 09:56:10	https://www.elmundo.es/economia/macroeconomia/2020/03/11/5e68b5a9fdddff77138b4658.html
2020-03-11 09:38:07	https://www.elmundo.es/tecnologia/videojuegos/2020/03/11/5e68b00efc6c83796c8b45b0.html
2020-03-11 09:31:45	https://www.elmundo.es/espana/2020/03/11/5e68aff0fdddfff8168b467e.html
2020-03-11 09:11:12	https://www.elmundo.es/economia/macroeconomia/2020/03/11/5e68ab07fdddffac5a8b4577.html
2020-03-11 08:58:27	https://www.elmundo.es/espana/2020/03/11/5e68a33521efa07a558b4630.html
2020-03-11 08:57:47	https://www.elmundo.es/economia/2020/03/11/5e68a6affdddff77138b4642.html
2020-03-11 08:21:37	https://www.elmundo.es/internacional/2020/03/11/5e689ee6fc6c833c108b4587.html
2020-03-11 06:59:07	https://www.elmundo.es/comunidad-valenciana/2020/03/11/5e688c2bfc6c83da618b458a.html

2020-03-11 06:33:53	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/11/5e68830bfc6c834f428b4578.html
2020-03-11 02:18:23	https://www.elmundo.es/deportes/2020/03/11/5e684a5dfdddff6c298b462e.html
2020-03-11 01:05:07	https://www.elmundo.es/opinion/2020/03/11/5e67fa7321efa0f16a8b460f.html
2020-03-11 01:01:34	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/11/5e67b7e621efa07a558b45e5.html
2020-03-11 00:57:24	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/11/5e67fc88fdddfff8168b464a.html
2020-03-11 00:35:22	https://www.elmundo.es/cultura/musica/2020/03/11/5e683236fc6c83b8438b4571.html
2020-03-10 23:00:23	https://www.elmundo.es/deportes/futbol/europa-league/2020/03/11/5e681bf7fc6c8329248b456e.html
2020-03-10 20:46:06	https://www.elmundo.es/madrid/2020/03/10/5e67fa54fdddffaa8e8b46e7.html
2020-03-10 20:03:38	https://www.elmundo.es/economia/2020/03/10/5e67f28521efa0f16a8b45fe.html
2020-03-10 19:23:32	https://www.elmundo.es/internacional/2020/03/10/5e67e91b21efa031788b462b.html
2020-03-10 18:25:54	https://www.elmundo.es/internacional/2020/03/10/5e67db7cfdddff4cbb8b4601.html
2020-03-10 17:13:47	https://www.elmundo.es/cultura/2020/03/10/5e67cabcfdddff4cbb8b45f7.html
2020-03-10 16:51:58	https://www.elmundo.es/espana/2020/03/10/5e67c2bd21efa008468b4608.html
2020-03-10 16:09:05	https://www.elmundo.es/deportes/futbol/primera-division/2020/03/10/5e67bb91fc6c834e638b469a.html
2020-03-10 15:09:52	https://www.elmundo.es/madrid/2020/03/10/5e67ab7dfdddfff7168b4625.html
2020-03-10 14:42:56	https://www.elmundo.es/cultura/cine/2020/03/10/5e67a75ffdddff4cbb8b45e5.html
2020-03-10 14:31:49	https://www.elmundo.es/economia/macroeconomia/2020/03/10/5e67a4bf21efa0087d8b45f3.html
2020-03-10 11:51:02	https://www.elmundo.es/economia/ahorro-y-consumo/2020/03/10/5e676337fc6c83304e8b4627.html
2020-03-10 11:40:47	https://www.elmundo.es/madrid/2020/03/10/5e677cb1fdddff4cbb8b45c4.html
2020-03-10 11:40:18	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/10/5e677bfdfdddff2b9e8b4577.html
2020-03-10 11:24:42	https://www.elmundo.es/espana/2020/03/10/5e6778e7fdddff851e8b4611.html

2020-03-10 10:00:20	https://www.elmundo.es/madrid/2020/03/10/5e67617821efa0950d8b45ef.html
2020-03-10 09:35:36	https://www.elmundo.es/internacional/2020/03/10/5e675e78fdddff43a98b45d1.html
2020-03-10 08:19:10	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/10/5e674d0efc6c83777d8b4604.html
2020-03-10 01:04:54	https://www.elmundo.es/tecnologia/2020/03/10/5e6660eefc6c8307608b45ea.html
2020-03-10 01:02:47	https://www.elmundo.es/economia/2020/03/10/5e66d835fc6c83d03c8b4624.html
2020-03-09 21:37:24	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/09/5e66b6ed21efa0ec188b471c.html
2020-03-09 20:45:21	https://www.elmundo.es/internacional/2020/03/09/5e66aac621efa026288b459f.html
2020-03-09 20:30:10	https://www.elmundo.es/deportes/baloncesto/2020/03/09/5e66a73921efa098188b47dc.html
2020-03-09 20:21:57	https://www.elmundo.es/deportes/futbol/primera-division/2020/03/09/5e66a432fc6c8302658b4606.html
2020-03-09 18:55:15	https://www.elmundo.es/madrid/2020/03/09/5e667296fc6c8397618b45de.html
2020-03-09 17:28:06	https://www.elmundo.es/internacional/2020/03/09/5e667994fdddff65468b460b.html
2020-03-09 12:30:56	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/09/5e6636ef21efa08c388b46cb.html
2020-03-09 09:56:34	https://www.elmundo.es/pais-vasco/2020/03/09/5e6612c021efa0082a8bf97d.html
2020-03-09 09:41:31	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/09/5e660f23fc6c8367138b45eb.html
2020-03-09 08:44:29	https://www.elmundo.es/espana/2020/03/09/5e6601dcfdddff1cb28b46d5.html
2020-03-09 08:17:42	https://www.elmundo.es/economia/macroeconomia/2020/03/09/5e65fb9721efa0c37c8b4634.html
2020-03-09 00:57:04	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/09/5e653bb021efa045038b4676.html
2020-03-08 21:47:01	https://www.elmundo.es/espana/2020/03/08/5e6567c421efa045038b4689.html
2020-03-08 17:23:03	https://www.elmundo.es/madrid/2020/03/08/5e6529e621efa08c388b4635.html
2020-03-08 14:11:17	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/08/5e64fbf2fc6c83621d8b4698.html

2020-03-08 12:11:24	https://www.elmundo.es/deportes/tenis/indian-wells/2020/03/08/5e64e09a21efa0ec188b45c2.html
2020-03-08 10:49:28	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/08/5e64cd53fc6c83ca228b4641.html
2020-03-08 10:18:57	https://www.elmundo.es/espana/2020/03/08/5e64c4cdfc6c834c698b45eb.html
2020-03-08 09:37:11	https://www.elmundo.es/deportes/formula-1/2020/03/08/5e64bcb7fdddff73698b4689.html
2020-03-07 19:12:39	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/07/5e63f21721efa01f428b4627.html
2020-03-07 13:36:56	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/07/5e63a2dffdddffeb038b467e.html
2020-03-07 13:01:52	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/07/5e6399e5fdddffd8768b4640.html
2020-03-07 11:16:03	https://www.elmundo.es/deportes/mas-deporte/2020/03/07/5e638250fc6c83b35f8b4606.html
2020-03-07 09:08:54	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/07/5e63606afdddff45b88b4572.html
2020-03-07 09:01:30	https://www.elmundo.es/deportes/baloncesto/nba/2020/03/07/5e6361c0fdddffcc328b46c2.html
2020-03-06 16:46:52	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/06/5e627db321efa0ab448b4588.html
2020-03-06 15:56:57	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/06/5e626a5a21efa0f9028b4597.html
2020-03-06 09:37:14	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/06/5e6219bdfc6c832f078b45e8.html
2020-03-06 08:18:00	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/06/5e620661fc6c837e7a8b45c3.html
2020-03-06 07:52:13	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/06/5e61fa17fc6c8345668b45d4.html
2020-03-06 07:40:41	https://www.elmundo.es/madrid/2020/03/06/5e61fe61fc6c83753e8b45c7.html
2020-03-05 20:48:34	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/05/5e616381fdddffa78a8b464e.html
2020-03-05 14:35:11	https://www.elmundo.es/baleares/ibiza/2020/03/05/5e610df721efa0e17b8b461d.html
2020-03-05 13:09:56	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/05/5e60f89ffc6c83c9058b4579.html
2020-03-05 10:09:45	https://www.elmundo.es/madrid/2020/03/05/5e60cfdcfc6c837e708b4623.html

2020-03-05 09:21:34	https://www.elmundo.es/economia/2020/03/05/5e60c2a121efa03c688b45e8.html
2020-03-05 08:48:00	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/05/5e60b30821efa031158b4570.html
2020-03-05 07:31:31	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/05/5e60aa17fc6c83b4288b4660.html
2020-03-05 06:29:20	https://www.elmundo.es/comunidad-valenciana/2020/03/05/5e5ffe76fdddff2a338b45ed.html
2020-03-04 16:24:50	https://www.elmundo.es/deportes/futbol/premier-league/2020/03/04/5e5fd63e21efa0e0368b458f.html
2020-03-04 10:26:57	https://www.elmundo.es/deportes/ciclismo/2020/03/04/5e5f825e21efa0b3458b46bb.html
2020-03-02 10:02:21	https://www.elmundo.es/ciencia-y-salud/salud/2020/03/02/5e5cd4ebfc6c83632e8b4644.html
2020-02-10 08:50:24	https://www.elmundo.es/economia/2020/02/10/5e41192bfdddffa4728b457b.html
2020-02-09 18:59:17	https://www.elmundo.es/economia/empresas/2020/02/09/5e40564cfdddff46438b4662.html
2020-02-09 09:22:36	https://www.elmundo.es/salud/2020/02/09/5e3fcf42fc6c8348638b462f.html
2020-02-08 07:13:41	https://www.elmundo.es/salud/2020/02/08/5e3e5d79fdddff3d748b45c6.html
2020-02-07 19:40:48	https://www.elmundo.es/ciencia-y-salud/salud/2020/02/07/5e3dbc97fc6c836e0d8b4631.html
2020-02-06 16:39:43	https://www.elmundo.es/salud/2020/02/06/5e3c4103fc6c8393788b45ac.html
2020-02-05 12:54:54	https://www.elmundo.es/salud/2020/02/05/5e3abb0221efa08f2a8b45a4.html
2020-02-04 16:26:37	https://www.elmundo.es/deportes/futbol/eurocopa/2020/02/04/5e399af6fc6c8341678b466c.html
2020-02-04 13:40:19	https://www.elmundo.es/ciencia-y-salud/salud/2020/02/04/5e397391fc6c83865c8b45a2.html
2020-02-04 09:52:11	https://www.elmundo.es/ciencia-y-salud/salud/2020/02/04/5e393e9cfc6c834d418b47c8.html
2020-02-02 20:45:39	https://www.elmundo.es/deportes/mas-deporte/2020/02/02/5e3734dffc6c8357498b45ae.html
2020-02-02 00:55:46	https://www.elmundo.es/ciencia-y-salud/salud/2020/02/02/5e35be7321efa074758b45ec.html
2020-01-31 13:57:42	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/31/5e343248fc6c8342178b45dd.html
2020-01-30 21:14:26	https://www.elmundo.es/deportes/mas-deporte/2020/01/30/5e32fa6efdddff9b138b457e.html

2020-01-29 20:46:05	https://www.elmundo.es/deportes/mas-deporte/2020/01/29/5e31eeb3fc6c831e1f8b462f.html
2020-01-29 18:18:12	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/29/5e31cb4521efa0962d8b45dd.html
2020-01-29 15:11:34	https://www.elmundo.es/deportes/futbol/2020/01/29/5e31a09521efa0983a8b45b1.html
2020-01-29 11:40:05	https://www.elmundo.es/salud/2020/01/29/5e316899fc6c83cb6a8b45e0.html
2020-01-29 01:25:27	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/29/5e30dee8fdddffa3608b45ba.html
2020-01-28 20:41:46	https://www.elmundo.es/salud/2020/01/28/5e309c3afdddff01428b4594.html
2020-01-28 16:13:16	https://www.elmundo.es/deportes/futbol/2020/01/28/5e305d8cfc6c83a86f8b45cf.html
2020-01-28 13:31:39	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/28/5e3037a921efa0413a8b4661.html
2020-01-28 10:32:46	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/28/5e300db421efa01f178b4609.html
2020-01-28 00:04:02	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/28/5e2f7a63fdddff9aa78b469b.html
2020-01-26 06:08:53	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/26/5e2d2cdafdddff8ba78b4574.html
2020-01-25 10:45:28	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/25/5e2c19e4fdddff444a8b4613.html
2020-01-21 14:41:06	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/21/5e270d00fdddffc9088b463f.html
2020-01-20 09:29:13	https://www.elmundo.es/ciencia-y-salud/salud/2020/01/20/5e25728b21efa0f4078b4693.html
2019-12-05 16:03:49	https://www.elmundo.es/ciencia-y-salud/salud/2019/12/05/5de63a10fdddffa8708b45bc.html
2019-10-10 00:08:34	https://www.elmundo.es/espana/2019/10/10/5d9e4ddafc6c83c86d8b4619.html
2019-10-01 11:16:51	https://www.elmundo.es/espana/2019/10/01/5d933442fdddffad598b45e9.html
2019-04-29 04:47:27	https://www.elmundo.es/espana/2019/04/29/5cc2ba57fdddff419f8b45e6.html
2019-04-28 23:48:51	https://www.elmundo.es/espana/2019/04/29/5cc618e3fc6c83441e8b456e.html
2019-04-28 22:20:58	https://www.elmundo.es/espana/2019/04/29/5cc610c6fdddff248e8b4615.html
2019-04-24 00:20:33	https://www.elmundo.es/opinion/2019/04/24/5cbf49c1fc6c8343038b4656.html

2019-04-18 00:53:03	https://www.elmundo.es/espana/2019/04/18/5cb7815efc6c835f3c8b45ab.html
2019-04-15 00:14:20	https://www.elmundo.es/espana/2019/04/15/5cb366d1fc6c83002e8b4580.html

Tabla D-16: Historial de noticias resultante para la temática de La COVID-19 en el periódico El Mundo

El País	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,43	0,38	0,35	0,38	0,19	0,43
Titular + Keywords	0,28	0,45	0,44	0,6	0,24	0,49
Titular + Resumen	0,22	0,46	0,36	0,5	0,21	0,46
Titular + Autor	0,35	0,34	0,34	0,33	0,23	0,36
Titular + Cuerpo	0,22	0,48	0,5	0,62	0,19	0,48
Keywords	0,21	0,54	0,38	0,56	0,21	0,55
Keywords + Resumen	0,19	0,53	0,37	0,48	0,18	0,54
Keywords + Autor	0,22	0,52	0,37	0,51	0,21	0,53
Keywords + Cuerpo	0,22	0,49	0,47	0,59	0,26	0,49
Resumen	0,12	0,3	0,23	0,3	0,1	0,31
Resumen + Autor	0,11	0,33	0,28	0,29	0,11	0,33
Resumen + Cuerpo	0,18	0,47	0,47	0,55	0,22	0,47
Autor	0,21	0,17	0,15	0,17	0,24	0,42
Autor + Cuerpo	0,18	0,47	0,5	0,57	0,24	0,47
Cuerpo	0,18	0,46	0,49	0,53	0,18	0,46

Tabla D-17: Media de resultados obtenidos por el periódico El País

El Mundo	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,25	0,6	0,62	0,55	0,19	0,6
Titular + Keywords	0,57	0,88	0,86	0,94	0,17	0,88
Titular + Resumen	0,3	0,6	0,61	0,55	0,23	0,6
Titular + Autor	0,23	0,66	0,59	0,66	0,15	0,66
Titular + Cuerpo	0,4	0,62	0,57	0,65	0,24	0,61
Keywords	0,7	0,84	0,84	0,96	0,13	0,91
Keywords + Resumen	0,66	0,83	0,57	0,9	0,2	0,85
Keywords + Autor	0,6	0,85	0,72	0,95	0,11	0,82
Keywords + Cuerpo	0,39	0,66	0,59	0,68	0,25	0,66
Resumen	0,25	0,17	0,12	0,17	0,29	0,25
Resumen + Autor	0,22	0,32	0,22	0,32	0,15	0,35
Resumen + Cuerpo	0,4	0,6	0,58	0,62	0,23	0,59
Autor	0,36	0,26	0,36	0,26	0,16	0,45
Autor + Cuerpo	0,37	0,61	0,59	0,63	0,23	0,61
Cuerpo	0,38	0,62	0,59	0,63	0,23	0,61

Tabla D-18: Media de resultados obtenidos por el periódico El Mundo

20 Minutos	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,54	0,57	0,54	0,54	0,23	0,6
Titular + Keywords	0,75	0,75	0,53	0,61	0,46	0,78
Titular + Resumen	0,54	0,52	0,45	0,49	0,32	0,52
Titular + Autor	0,54	0,6	0,54	0,57	0,26	0,6
Titular + Cuerpo	0,32	0,5	0,4	0,66	0,2	0,5
Keywords	0,75	0,66	0,69	0,6	0,54	0,66
Keywords + Resumen	0,61	0,69	0,66	0,49	0,36	0,69
Keywords + Autor	0,78	0,66	0,69	0,57	0,54	0,66
Keywords + Cuerpo	0,38	0,53	0,44	0,68	0,23	0,53
Resumen	0,29	0,46	0,51	0,46	0,36	0,46
Resumen + Autor	0,31	0,4	0,45	0,39	0,26	0,43
Resumen + Cuerpo	0,35	0,5	0,43	0,66	0,2	0,5
Autor	0,24	0,17	0,17	0,17	0,27	0,51
Autor + Cuerpo	0,35	0,53	0,37	0,63	0,2	0,53
Cuerpo	0,35	0,5	0,37	0,63	0,2	0,5

Tabla D-19: Media de resultados obtenidos por el periódico 20 Minutos

El Confidencial	Similitud Coseno Vectorización Word2Vec Estudio Simple	Similitud Jaccard Estudio Simple	Similitud Coseno Vectorización tf-idf	Similitud Coseno Vectorización BOW	Similitud Coseno Vectorización Word2Vec Estudio por horas	Similitud Jaccard Estudio por horas
Atributo	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**	**f1-Score**
Titular	0,34	0,25	0,36	0,25	0,22	0,22
Titular + Keywords	0,45	0,82	0,36	0,36	0,15	0,81
Titular + Resumen	0,23	0,31	0,36	0,28	0,15	0,35
Titular + Autor	0,4	0,26	0,39	0,26	0,25	0,19
Titular + Cuerpo	0,36	0,28	0,36	0,21	0,29	0,28
Keywords	0,48	0,83	0,43	0,57	0,15	0,82
Keywords + Resumen	0,35	0,78	0,44	0,6	0,18	0,76
Keywords + Autor	0,59	0,81	0,36	0,57	0,15	0,78
Keywords + Cuerpo	0,42	0,42	0,42	0,23	0,31	0,39
Resumen	0,19	0,35	0,5	0,35	0,15	0,32
Resumen + Autor	0,22	0,31	0,5	0,34	0,15	0,29
Resumen + Cuerpo	0,31	0,3	0,38	0,22	0,31	0,26
Autor	0,18	0,4	0,52	0,4	0,15	0,39
Autor + Cuerpo	0,38	0,27	0,35	0,22	0,29	0,25
Cuerpo	0,33	0,26	0,35	0,21	0,29	0,26

Tabla D-20: Media de resultados obtenidos por el periódico El Confidencial